Using Document Clustering and Language Modelling in Mediated Information Retrieval

Gheorghe Muresan

A thesis submitted in partial fulfilment of the
requirements of
The Robert Gordon University
for the degree of Doctor of Philosophy

January 2002

Collaborating establishment, Ubilab, Union Bank of Switzerland, Zurich.

Acknowledgments

I am most grateful to Prof. David Harper, my director of studies, who introduced me to Information Retrieval and guided my first steps into research. He was always challenging but supportive. In the latter stage of my PhD he helped me focus on "the story to tell" rather than continue to investigate all minute detail that may or may not be relevant to my investigation. He was patient to read again and again my "too long sentences".

I am also indebted to several supervisors and advisors that I have had along the way, who advised me on various aspects of my work. John Boyle and Peter Lowit suggested ideas on the user interface design; Ayse Goker contributed ideas towards the design of user experiments and suggested improvements to the introductory chapter of my thesis; Mourad Mechkour worked with me and took the initial lead in the software design.

I was lucky enough to have inspiring colleagues and friends. When I started, Jan-Jaap IJdens was only too happy to answer my questions about IR, Unix and anything else. Dave Hendry introduced me to design patterns, which have become the basis of all my software. Joemon Jose, whose trail I followed, was a model of dedication to research. More recently, discussions with Angus MacLean, Daqing He and Bill Teahan helped me clarify my ideas. Many thanks to them and to Gerrit Renker for reading and commenting my papers and parts of my thesis. Gareth Palmer, Malcolm Souter and Beatriz Garmendia-Doval's contributions to our social life are also to be thanked.

This research was generously sponsored by Ubilab, the Union Bank of Switzerland, through Prof. Hans-Peter Frei. Ubilab were not only my sponsors, but also collaborators and advisers. I learnt a lot from working with and talking to Tore Bratvold, Marisa Barja, Jussi Myllymaeki, Gabriele Sonnenberger and Matthew Chalmers.

I also wish to thank various organisations that sponsored my participation in various conferences: CID for RIAO'97, ACM for SIGIR'99 and SIGIR'00; CEPIS for BCS-IRSG'98, BCS-IRSG'99, BCS-IRSG'00 and the final MIRA conference; ERCIM for ECDL'01.

Moreover, I would like to thank Alex Wilson, Acting Head, and Prof. Susan Craw, Head of School, who continued to support me financially after my sponsorship ended. Thanks should be extended to the administrative staff in the School and to the technical support team, especially John Riddoch, who were always quick to solve my problems.

Emma Forster and Rhona Gibson, as University Research Support Officers were very helpful in all official matters. I have benefited a lot from the research training courses they organised.

I am lucky to have lived in Scotland while conducting this difficult research work. The beautiful Bens and Glens have helped alleviate the stress. Also thanks to many people in the hill-walking, running, tennis, squash and ski clubs who helped me stay healthy. The enthusiasm of the fellow members of the Research Student Association, Chetna Patel, John Little, Daniel Aklil, David McMinn and others have created an excellent environment for exchanging ideas and socialising.

Many thanks to my family, my mother Margit, my brothers Luci and Cristi, and especially to my wife Elisa, who often had to cope with my stress and bad mood. They were always loving, understanding and supportive.

Finally, I dedicate this thesis to the memory of my father who was my role model and from whom I learnt to never waste a minute.

Gheorghe Muresan
January
2002

Abstract

Our work addresses a well documented problem: users are frequently unable to articulate a query that clearly and comprehensively expresses their information need. This can be attributed to the information need being too ambiguous and not clearly defined in the user's mind, to a lack of knowledge of the domain of interest on the part of the user, to a lack of understanding of a retrieval system's conceptual model, or to an inability to use a certain query syntax.

This thesis proposes a software tool that emulates the human search mediator. It helps a user explore a domain of interest, learn its structure, terminology and key concepts, and clarify and refine an information need. It can also help a user generate high-quality queries for searching the World Wide Web or other such large and heterogeneous document collections.

Our work was inspired by library studies which have highlighted the role of the librarian in helping the user explore her information need, define the problem to be solved, articulate a formulation of the information need and adapt it for the retrieval system at hand in order to get information.

Our approach, **mediated access through a clustered collection**, is based on an information access environment in which the user can explore a relatively small, well structured, pre-clustered document collection covering a particular subject domain, in order to understand the concepts encompassed and to clarify and refine her information need. At the same time, the user can ostensively indicate clusters and documents of interest so that the system builds a model of the user's topic of interest. Based on this model, the system assists and guides the user's exploration, or generates 'mediated queries' that can be used to search other collections.

We present the design and evaluation of WebCluster, a system that reifies the concept of mediated retrieval. Additionally, a variety of mediation experiments are presented,

which provide guidelines as to which mediation strategies are more appropriate for different types of tasks.

A set of experiments is presented that evaluate document clustering's capacity to group together topical documents and support mediation. In this context we propose and experimentally test a new formulation for the cluster hypothesis.

We also look at the ability of language models to convey content, to represent topics and to highlight specific concepts in a given context. They are also successfully applied to generate flexible, task-dependent cluster representatives for supporting exploration through browsing and respectively searching.

Our experimental results show that mediation has potential to significantly improve user queries and consequently the retrieval effectiveness.

Contents

Chapter 1

Introduction

1.1 Problem statement

We live by information, not by sight.

- Baltasar Gracian, *The Art of Worldly Wisdom* -

A half century of pioneering concepts and fundamental research have been digitized and indexed in a variety of ways in this special collection of works published by ACM since its inception. The ACM Digital Library includes bibliographic information, abstracts, reviews, and full texts.

- The ACM Digital Library, *http://portal.acm.org* -

When Vannevar Bush envisioned his 'memex' in 1945, at the beginning of the modern information age, he dreamed of tackling the problem of information overload at that time. Today, with the advent of more and more powerful computers, we are still struggling to cope with the vast amount of information that bombards us.

People need information to solve problems or to get informed. They may want something simple but necessary, such as a train timetable, or they may want to better understand the current situation or the origins of the Middle East conflict. They may want information for private entertainment or for work.

Most of the time, all the information that a user needs is freely available in electronic form in a digital library, in an institution-wide intranet, or on the World-Wide Web.

However, it is not readily accessible. It needs to be found. And the larger the amount of information that becomes available, the more difficult it is to locate the specific information needed at one point.

Studies of users searching for information in electronic document collections, including the Web, show that unassisted online searching is difficult for end users, with a failure rate close to 50%. "Before launching a search, take the time to think about what unique words or phrases are likely to appear in the information you want to find and try them first" is one of the online tips for "highly effective web searching"[1]. However, users have difficulties both in choosing search terms to represent their problem and in re-formulating their query in case of failure [Nor99].

This is the very issue this thesis addresses: the users' inability to generate high-quality queries, which clearly and comprehensively convey their information need. This introductory chapter explores in more detail some causes for this situation, looks at some approaches that have tried to address it, and explains the basic idea of **mediated retrieval**, our proposed solution to tackle it.

1.2 Information need formulation

1.2.1 Problems with query formulation

Early information retrieval systems (IRS) were mainly used by information specialists. This was reflected in their design and development, and in the research aimed at improving them. The approach taken was *systemic*: users were considered competent enough to formulate queries that accurately described their information need, so the main research effort was in developing indexing models as well as data structures and search algorithms that produced good efficiency and effectiveness of retrieval. Even the design of the test collections and of the experiments to evaluate retrieval was influenced by this approach: the searches were in batch mode, based on a fixed set of test queries and corresponding relevance judgements, and the evaluation consisted in estimating the quality of the retrieved set of documents.

[1] http://beta.peachpit.com

With the increased availability of computers and of searching tools for a wider audience, it has become clear that these systems are difficult to use by un-trained users. User studies, including analyses of Web search engine logs [JSBS98], indicate a high rate of search failures and provide some evidence as to why that is the case. Users often display:

- little knowledge of the appropriate vocabulary (especially for specialised domains, with specific terminology).
- inability to use advanced query language syntax, such as Boolean operators.
- false impression that "the computer knows what I want" i.e. a mis-placed expectation that the system is aware of the user's context.
- lack of a search strategy. Most search sessions consist of few queries and display no real exploration of the information space. Query re-formulation is rarely applied.
- lack of a clear understanding of the system's conceptual model. Rather than providing precise queries, containing terms with high power of discrimination, users tend to submit very short queries, often made up of ambiguous or common words.

Let us look a bit deeper at why users' queries are so often ineffective for retrieval. Prior to allowing searching, information retrieval systems *index* the collection of documents to be searched, by associating sets of index terms to each document and applying a *weighting model* that estimates the contribution of each term to each document. The selection of terms to represent each document and the weighting scheme applied are intended to achieve a trade-off between *representation* (indicating what the document is about) and *discrimination* (indicating how the document is different from other documents in the collection). When doing a search based on a query, the system compares the query with the document representations obtained through indexing. It is obvious that a good understanding of the indexing (and searching) model increases the user's chances of producing a good query. In order to generate a high quality query, a user would also need to know the vocabulary or the terminology of the collection and the distribution of the terms over the set of document representatives. This is made explicit by Ponte and Croft: "A user that understands our model will tend to think in terms of which words will help the system distinguish the documents of interest from everything else. We feel that

if we can get users to think in this manner they will be able to formulate queries that will better express their information needs in a manner useful to the retrieval system" [PC98].

The importance of the query quality is confirmed by the Text REtrieval Conference (TREC)[2] series of experiments: in simulated retrieval sessions in which the description of the user's topic was rich and clearly specified, the difference of performance between participating retrieval systems was minimal [VH00], suggesting that the generation of good queries is more important than the improvement of weighting schemes or search algorithms.

Most information seekers are not trained to think in terms of IR models and search strategies, and do not consider the underlying indexing and searching models when formulating their query. Hence, they cannot be expected to produce good queries unassisted. This is exactly the issue that we are trying to address by designing a system that can assist the user in generating high-quality queries, in order to improve the search effectiveness.

1.2.2 Cognitive aspects of the information seeking process

The discussion of the query formulation problem, in the previous subsection, assumed that the user knows exactly what she wants, and that all she needs is support in better conveying her information need. However, it is often the case that the user does an exploratory search, when she does not quite know what she wants ("I can't say what I want, but I will recognise it when I see it") and one important part of the information seeking process is the user's information need clarification and refinement [Bel00].

One of the first researchers to investigate the cognitive aspects of the information seeking process was Taylor, who looked at the search process as a description of an area of doubt in which the question is open-ended, negotiable, and dynamic [Tay68]. Taylor identified four levels of the information need, as it appears and then evolves in the user's mind during the interaction with an intermediary searcher:

1. the *visceral* need - the actual, but unexpressed need for information. It may be only a vague sort of dissatisfaction, probably inexpressible in linguistic terms.

[2] http://trec.nist.gov/

2. the *conscious* need - a mental description of an ill-defined area of indecision. It may be an ambiguous and rambling statement.

3. the *formalized* need - a qualified and rational statement of the user's question. It is a description in concrete terms of the area of doubt which may or may not take into consideration the constraints of the retrieval system.

4. the *compromised* need - the query as presented to the information system. It is a recast in anticipation of what the system can deliver and it is dependent on the systems's domain coverage, specialisation, indexing model, and format.

The information specialist mediating the search is particularly useful in the final stage of the information need formulation, corresponding to the query generation, as she is (expected to be) knowledgeable with regards to the particularities of the information system employed and also with the structure and terminology of the domain explored. However, by being a communication partner to the user, and by eliciting the user's problem and context, she can help the user through all the stages of the search process.

Taylor also identified five filters employed by librarians to select significant data in order to assist the user:

1. determination of the subject. Determine the general delineation of the subject, its limits and structure.

2. objective and motivation. Further qualify the subject, ascertaining details such as size, shape and form of possible answers.

3. personal characteristics of the inquirer. Establish the context of the enquiry.

4. relationship of enquiry description to file or system organization. Formulate the query in the system's terms.

5. anticipated or acceptable answers. Process feedback and re-formulate query.

Taylor's work was extended and refined by Belkin and his colleagues [BOB82], who challenged the assumptions of the traditional *best-match* model underlying the design of most IR systems at the time. They stated that it was the exception, rather than the rule,

that a user would be able to accurately describe her information need.

Belkin's *anomalous state of knowledge (ASK)* hypothesis proposed that an information need typically arises from an anomaly in the user's state of knowledge concerning some topic or situation and that the user is unable to specify precisely what is needed to solve that anomaly. The information need is usually not a need in itself, but rather a means towards the resolution of a goal or a problem. Faced with the problem, the user realizes that her state of knowledge is inadequate for solving it, so she proceeds to explore the problem domain in order to gain a better understand of it and of the information that is needed in order to solve the problem.

This hypothesis raises an extra challenge for our endeavour: we cannot assume that the user's information need is clear, and expressing it in query form is all that she needs help with. We need to consider not only the *verificative/analytic* aspect of the information seeking process, when the user searches for some information known to exist, but also the *explorative* aspect, when the user needs to be supported in exploring her problem domain and in clarifying and refining her information need.

1.3 Previous approaches to information seeking support

If I have seen further it is by standing on shoulders of giants.

- Sir Isaac Newton -

The previous section suggests the two complementary issues that we intend to address in this thesis:

1. support for the user in exploring a problem domain, and in clarifying and refining her information need.
2. support for the user in generating high-quality queries.

Before presenting our own proposed solution, mediated retrieval, we are looking at research that has inspired, and that supports our work. No comprehensive review of this research is intended; we only concentrate on work that closely relates to our approach:

1. highly interactive retrieval systems that employ information visualization for supporting exploration and concept formation.

2. tools and techniques for query formulation or expansion.

These approaches are discussed in the following subsections. A further subsection is dedicated to a complementary technique, *relevance feedback (RF)*, which attempts to interpret the user's response to the information seen, in order to improve the system's model of the user's information need. A full subsection is devoted to relevance feedback as its conceptual model is similar to that of our approach.

1.3.1 Information visualization for interactive exploration of the information space

Belkin's cognitive model suggests the need for designing iterative and highly interactive retrieval systems that support the user's exploration of a topic of interest in order to solve a problem. Rather than expecting precise queries from the user and attempting to answer them, such systems expect a general statement of the problem. They then infer the knowledge structures underlying the information need and offer relevant information expected to enhance the user's understanding of the problem domain, and the refinement of the problem specification. It is expected that during the interaction with the system the user's understanding of the problem evolves, various aspects of the problem become apparent, and the user's actual information need gradually becomes clearer, more refined, and easier to articulate. The user either finds the solution to her problem during the interaction, or reaches a level of understanding of the topic that affords an accurate and comprehensive description or formulation of the information need or of the expected solution. In the latter case, the system's operating mode should change to high precision best-match searching, in order to identify documents that correspond to the information need description.

Belkin argued strongly in favour of considering the *user* the central component of an IR system, and the *interaction* the central process of IR [Bel93]. He incorporated these ideas in the design of BRAQUE, an IR system that allows the user to directly interact with text and manipulate text, and to change her behaviour and strategy in response

to that interaction with information [BMC93]. Oddy's THOMAS [Odd77] was another system that made the user's interaction the central process of retrieval; rather than allowing query formulation, it guided the user's exploration by indicating associations between documents, subjects and authors. A more flexible system was Croft and Thompson's I^3R system [Odd77], which allowed the user to combine browsing and searching by specification, and encouraged the user to participate actively in the retrieval process, by eliciting and incorporating the user's knowledge in the representation process. Bates allowed for the dynamic evolution of the user's information need during and due to the interaction with information; her 'berrypicking' model assumed that users usually collect useful information and gradually build a solution for their problem rather than suddenly arriving at the solution somewhere in the information space [Bat89].

The implementation of such interactive systems was facilitated by the advent of graphical user interfaces that afford the exploration of various sources of information. Such interfaces support *concept formation*: during exploration, a mental model of the domain or topic of interest develops in the user's mind. A cognitive map, or rather a cognitive collage [Spe00] is built based on the documents or pieces of information that the user sees and interprets while browsing the information space. There are two determinants of the browsing strategy: a *cognitive* one, supporting a consciously planned browsing strategy, based on the analysis and interpretation of information, as well as on relationships between documents and structural clues, and a *perceptual* determinant, supporting a less well organised strategy, where the user acts based on what she's seen at one point, hoping to find relevant documents by serendipity.

Typically, these exploration tools offer support for navigation by:

- indicating relationships between documents in the information space, such as topical or semantic structure. Such artefacts are: **cone trees** [RMC91, Hea97], **tree maps** [Shn92][3], **hyperbolic trees** [LRP95, MJS+97], **themescapes**[4], **workscapes**[5], or **Kohonen self-organised maps (SOM)** [HF99].

- indicating relationships between documents of the information space and a specified

[3] http://www.cs.umd.edu/hcil/treemap3/
[4] http://www.cartia.com
[5] http://www.maya.com/

query or topic. **Tilebars** [Hea95], **InfoCrystal** [Spo94], and **Vibe** [Kor91] are such examples.

- using filters based on meta-information (date, title, popularity). They are illustrated in interfaces such as **HomeFinder** [WS92] or **FilmFinder** [AS94].

What these approaches to visualization have in common is that they use various metaphors or organisation schemes to support the exploration of a set of documents - either of the whole information space for a certain application [Kor91], or of a subset deemed to contain the solution to the user's information need. It is also possible to combine a variety of tools in order to offer complex assistance to the users in the form of visualization of classified collections and of ontologies for the user's domain of interest [Pol97b, Pol97a, Pra99].

A combination of *searching* and *browsing* strategies is usually supported. The former tend to be more formal, analytical, goal-driven, deterministic, and typically employed by experienced searchers. The latter tend to be more informal, opportunistic, heuristic, dependent on interaction and on the interpretation of the information found so far. They are more typically employed by information seekers with less searching experience, but possibly with more domain knowledge [Mar95].

While analytical strategies fare well with people who know what they are looking for, browsing is especially useful for searchers who need to first clarify their information need, as it helps the user

- gain an overview of the information space (the domain of interest), of its terminology, topics and concepts, and of the relationships between topics (the structure of the domain).

- reduce the cognitive load, as the human is required to recognize relevant information, rather than to explicitly formulate a request.

- to serendipitously discover useful information in unexpected places and learn more about the domain by gaining insight into unsuspected associations between topics.

- develop a model of the relevant documents, of their representation, and a formal strategy for further searching.

Some systems do not support querying at all, and rely entirely on browsing for supporting the user's exploration. These query-less systems are more common for cases when the media of the documents does not afford an easy formulation of an information need, and rely instead on the user recognising useful items or features. All the documents may be organised (clustered, for example) and presented to the user, or just a sample of representative ones. The user can explore the documents and request "more like this", as is the case in Campbell's 'ostensive model' [Cam95, CR96]. The system relies on a measure of document-document similarity to satisfy the user, rather than on a potentially imprecise query. This approach has the advantage that documents are in the same space, so modelling their similarity has a stronger theoretical support than modelling query-document similarity [BSR93].

The idea of query-less systems has obvious limitations: it is not practical for huge collections such as the Web. Moreover, such systems are not effective if the user knows exactly what she wants and would prefer to get access to the relevant documents directly, rather than following a potentially lengthy browsing procedure.

Another model for IR interaction, supported by some experimental systems, is derived from *foraging theories* [PC95, Cha00, PCW01]. Some researchers have shown that the human's behaviour when searching for information is similar to the behaviour of the animals, and of the hunter-gatherers, looking for food, hence the name "informavores" [Cha00]. Useful information, like food or other resource, is not evenly distributed, but patchy. Firstly, the forager needs to find a good patch. Then, the longer a patch is exploited, the lower the returns will be, until the patch is over-grazed and worthless. However, the time spent searching for a new patch is unprofitable. The *marginal value theorem* gives the optimal strategy for maximizing benefits per costs: move from a patch when the rate of return falls below the average return rate over the whole region. This approach also reflects Marcia Bates's "berry-picking" strategy of solving a task by collecting relevant information from here and there, rather than trying to find a source of information that solves the task[Bat89].

1.3.2 Query formulation support

The techniques described here have been designed to support the user's query formulation, typically in situations when the user is aware of a certain information need, but has problems in formulating a good query. The reason may be a low level of familiarity with the domain investigated and its terminology, or simply with the distribution of terms over the documents of the collection investigated.

1. **Aid-word list**

   According to research in Cognitive Psychology, it is easier to recognize terms that describe an information need than to generate them. An aid-word list, suggesting possible query terms to the user, can be effective particularly when a controlled, small-size vocabulary was used for indexing. A somewhat less effective approach is the spell-checker.

2. **Thesauri**

   Thesauri convey not only the vocabulary of the domain, but also relationships (equivalence, generalisation, specialisation) between terms. The alternative terms "ontology" or "taxonomy" can be used especially when the thesaurus captures, in a hierarchic structure, the relationships between the concepts of a specialised domain.

   Typically, they are used as sources of alternative or additional terms to the query terms supplied by the user in order to increase the precision or the coverage of the query. Building thesauri can be done manually, by human domain experts, or automatically, based on collocation of terms in documents (term clustering) [SJ71]. The use of thesauri for query expansion can be done either explicitly, by offering the user the choice of alternative (or extra) terms for the query formulation, or implicitly, by automatically replacing the terms of the query with all the words in their family.

3. **Automatic local analysis**

   This class of techniques is somewhat similar to relevance feedback in that it uses information from documents retrieved following an initial query in order to reformulate the query. The difference is that no relevance judgements are used. Instead, the retrieved set is seen as a context in which relationships between terms can be

observed and modelled [BYRN99, XC96]. The query is expanded by adding terms estimated to refer to the same topic.

4. **Automatic global analysis**

   This class of techniques uses the whole document collection to build a thesaurus either by considering the similarity between terms and building sets of nearest neighbours terms, or by applying (hierarchic) document clustering and considering the representatives of bottom clusters as topic representatives. Then, terms for expansion are considered based on their similarity to the query as a whole rather than on their similarity to individual query terms [QF93]. Similar approaches expand the query based on term similarity obtained through term clustering [SJ71].

5. **Latent semantic analysis (LSA)**

   This technique, also known as *latent semantic indexing (LSI)* [DFK88, DDF+90, Fol96] attempts to address the problem of word synonymy and polysemy. LSI uses *singular value decomposition*, a powerful and fully automatic statistical method, to uncover the associations among terms in a large collection of texts, to create a virtual *concept space*, and to exploit it to improve retrieval. For example, by analysing a collection of texts, LSI will learn that "laptop" and "portable" occur in many of the same contexts, and that queries about one should probably retrieve documents about the other. LSI techniques are particularly useful when high recall is necessary, when text descriptions are short, when user input or texts are noisy, or when there is a need to retrieve information in multiple languages without requiring translation of queries or documents. LSI produces an indexing of collection documents based on virtual concepts, rather than terms, and also converts a query from a set of terms into a set of concepts.

6. **Personalization**

   It is difficult for a system to derive the user's information need based on a short or vague query, and to retrieve the right documents. Based on a set of queries, though, or even on all the queries of a set of search sessions, the system can use Machine Learning (ML) techniques to build a *user profile*. Subsequently, during future searches, the system can use the profile, or the *user context* to expand or

disambiguate the query, in order to improve the retrieval effectiveness.

A difficulty of this endeavour is the fact that the user may be interested in and may conduct searches related to more than one topic. Moreover, the user's interests may appear, evolve and disappear over time, at different rates. The difficulty is exacerbated on the Web, where the identification of the user is hampered by both technical and privacy issues [GH01, HGH01].

1.3.3 Relevance feedback (RF)

In an interactive retrieval system a query can often be improved (and re-submitted) based on the relevance judgements that the user makes on retrieved and examined documents. Following an initial, tentative user query, the system returns a set of documents estimated to be relevant to the user's information need. The user can explore these documents and mark them as relevant or non-relevant. Based on the user's actions, the system builds a model of the topic(s) the user is interested in. Most often, the *topic model* is represented as a set or ranked list of terms that are typical for the documents marked relevant and atypical for the documents marked non-relevant. Weights may be assigned to terms in order to indicate their contribution to the topic representation.

There are two distinct ways to subsequently use this topic model:

Automatic relevance feedback implies that the original query is automatically expanded with 'good' terms (and possibly re-weighted), and is re-submitted to the system without user consultation.

This type of relevance feedback may be not quite intuitive for users, and the results of applying it may seem confusing, especially in the case of negative feedback [Dun97]. When the user indicates that a document is not relevant, what does that mean for a complex, multi-topical information need, or for a topic with multiple *aspects* ? Does it mean that the topic is wrong or that the aspect of the topic is not the right one ?

Even for positive feedback, there is potential for producing poor query expansion terms. The sample of relevant documents may be small, terms may be extracted from non-relevant sections of relevant documents, and some relevant terms may not

be good discriminators and may also attract non-relevant topics. However, it has been shown that automatic relevance feedback does improve retrieval performance [SB90].

Interactive relevance feedback implies that the system proposes to the user a set of terms which appear to be typical for the documents judged relevant. It is the user who chooses to accept or reject the system's recommended query terms before the new query is submitted for a new search.

A problem with this form of relevance feedback is its limited ability to capture and represent important aspects of what makes a document relevant to a query, such as particular term co-occurrence patterns, discourse structure or style, because the users typically select single words, rather than phrases or groups of words [PCS00]. On the other hand, this approach has the advantage that it allows more capable users to concentrate on the aspects that are relevant for them.

The relevance feedback process can be applied iteratively until the search results are satisfactory.

Koenemann and Belkin have shown that interactive relevance feedback can significantly improve performance even for novice searchers, who make relatively few relevance judgements [KB96]. They compared four systems: a *baseline* system that allowed the user to manually re-formulate the query, following the examination of the search results, and three experimental systems based on relevance feedback that allowed the users progressively increased access to the RF mechanisms. An *opaque* system treated relevance feedback as a black-box, hiding its functionality from the user[6]; a *transparent* system showed the user the terms added to the query based on the user's relevance judgements; and a *penetrable* system allowed the user to accept or reject the terms proposed by the system. All three RF-based system performed significantly better than the baseline system, with small improvements for the systems with increased transparency. The subjects preferred the penetrable system that allowed them to understand and control how relevance feedback adjusted the query.

[6]Their so-called opaque system implements automatic relevance feedback.

While an improvement in retrieval effectiveness through interactive query expansion was observed by various other researchers [Eft00], there have also been experiments that were unsuccessful in that respect. Some indicated that users may ignore the RF functionality or get confused by it [BDPJ97], or may simply not recognise good terms proposed by the system, presumably because they are not familiar with the vocabulary or with the distribution of words in the documents [MR97, HMM99].

After comparing the conditions of the two sets of experiments, successful and unsuccessful, we suggest some conclusions which we believe are also valid for our mediated approach to retrieval:

- In order for the searchers to use the RF component, they need to know about it, to understand it, and to either like it and trust it (from previous use, or from a convincing tutorial) or to be constrained to use it (through a clever design of the user interface).

- In order for interactive RF to be useful, users need to have a minimum understanding of the underlying indexing and searching model and of the domain vocabulary.

- The experimental conditions and perhaps the motivation of the searchers may be important factors in the outcome of user experiments. Firstly, humans tend to behave differently under experimental conditions (the so-called Hawthorne effect), compared to 'normal' conditions, presumably under the influence of the perceived expectations. Secondly, students in Information Science are probably more likely to be interested in the internals of the search process and more eager to be in control of the system and of the retrieval process than other types of subjects. Especially if encouraged by course credits, they are also probably more likely to accurately follow the experimental procedure or assigned search strategy and to use the functions of the system as requested in the instructions. Therefore they may behave less naturally than the 'average information seeker', so caution should be exercised in generalising behavioural observations. This may be a possible explanation for the lack of correlation between, on the one hand, Koenemann and Belkin's observation [KB96], backed up by others such as Bates [Bat90], that users want *control*, and on the other hand, results showing that users are task-oriented [MDKL93], that they

ignore functions of the system, that they mis-use functions of the system, and that they want 'magic', without interest on how that happens [Cro95].

The implication for the design of IR systems is that flexibility is necessary, so that users can choose between *control* and *magic*.

Note that relevance feedback can also be produced based on annotated paragraphs or marked words [GPS98]. The effectiveness of this method can be improved if it is combined with *visualization tools* that indicate the relationship between the query and the retrieved documents, showing the contribution of the query terms to the retrieval decision, as in TileBars [Hea95].

In order to avoid the user interface complexity and the need for user cooperation introduced by explicit relevance feedback, alternative approaches have been proposed. One such approach, still under investigation, is *implicit* relevance feedback: the user's behaviour and actions, such as reading time, scrolling, request for a summary or other such interactions with documents are interpreted in order to derive the user's interest in the documents retrieved or shown by the system. The results of such research, although still inconclusive, look promising [WJR01, KB01].

An alternative which has already proven to significant improve retrieval effectiveness is "blind" (or pseudo-) relevance feedback, which ignores completely the user's behaviour or reaction to the documents seen [All95, RWB00]. It consists of assuming that the top ranked documents following a best-match search are relevant and of re-formulating the query accordingly, before a new search is done. Blind negative feedback can also be applied by assuming that the bottom documents retrieved by a best-match search are not relevant.

1.4 The WebCluster approach to mediated retrieval

Various studies have shown that mediated searches have a far higher success rate than un-mediated ones [Nor96]. This can be attributed to the interaction between the end-user and the mediator, typically a librarian or a professional intermediary searcher. During the interaction, the mediator elicits information from the user, in order to establish the

context of the investigation, the various aspects of the information need and the level of detail or abstraction required. The mediator thus helps the user analyse and explore the domain of her investigation as well as articulate and refine her information need. If necessary, after the information need is clarified, the mediator also formulates the appropriate query for the retrieval system at hand. In a library, the librarian can also take the user to the shelves that broadly cover the domain investigated and can indicate starting points for browsing the topic of interest.

We propose emulating the mediator's role by designing a system to support the user during a search session. It will interact with the user and will help her clarify and refine her information need; it will then build the right query or set of queries and will retrieve the matching documents in an attempt to satisfy the user's need.

Studies of mediated and un-mediated searches (such as [Nor96, SGRM96, Spi97, SSW97, SGR98, Nor99]) give an indication of the role of the mediator in the search process (information need formulation, knowledge of the information space structure, search strategy, query reformulation, ...). This thesis explores aspects of a system-mediated process and tries to assess how successfully it can replace or support the human mediator in improving the user's retrieval process and, ultimately, retrieval effectiveness.

The name of our proof-of-concept system, **WebCluster**, indicates the domain that initially suggested the idea of this research, and that was targeted in the first application of the mediated access concept: the World Wide Web. The Web epitomizes the situation that this thesis addresses. It is a huge, dynamic and largely unstructured document collection (the hyperlinks produce some semantic structure, but this is rather local and limited to the webpage authors' knowledge). It is widely accessible, so no expertise in search strategies and query languages or an understanding of indexing and retrieval models can be expected from users. However, as the thesis attempts to show, the results and ideas presented here are not restricted to searching the Web. They can be applied to information seeking on any large and heterogeneous collection which lacks structural organization and affords query-based searching as the sole means for exploration.

In designing WebCluster, we combined the two approaches, *cognitive* and *systemic*. Our main target is the user framed by Belkin's ASK model, who has a problem to solve, but does not know what information is needed or how that information could be obtained. We therefore supply a highly interactive interface that allows the user to explore the so-called 'source collection', a specialised collection chosen so that it is representative for her problem domain. During this exploration, the user becomes familiar with the terminology and topical structure of the domain of interest, and may find sufficient relevant documents to solve her problem. She has the option to mark documents or clusters of documents that are relevant for her problem in the sense of either clarifying some aspects, or suggesting solutions to the problem. Based on this interaction and specifically on the user's actions, the system 'learns' the problem's structure, context and boundaries, and it effectively builds a *topic model* for it. Subsequently, the behaviour of WebCluster is systemic: based on the topic model and on its internal indexing model and syntactic rules, it builds an optimal query that accurately and comprehensively reflects the user's topic of interest and performs a search on the Web or any other 'target collection' likely to contain answers or solutions to the user's problem.

Most often in the literature *mediation* means the involvement of a human intermediary who interacts with the user (asks questions, prompts various actions, ...) and in doing so assists the user in formulating, refining, and hopefully solving an information need. *Assistance* is the term usually used to denote the support a system gives a user towards a certain goal. However, WebCluster's intended functionality matches so closely that of the human mediator, that we preferred the term 'mediation' for its function.

Apart from the main function of WebCluster, sketched above and analysed in more detail in the rest of this thesis, other uses can also be envisaged. Firstly, novice computer users, such as some patrons in public libraries, may prefer to talk to a librarian instead of a machine. WebCluster can act as a librarian's tool by supporting the exploration of the user problem's domain and also a good query formulation. Secondly, even people who know exactly what they want may have problems in formulating their information need, or may be dissatisfied with the search results based on their queries. They can be supported in formulating better queries in order to improve their search results.

1.4.1 Conclusions

We propose the concept of *system-based mediation* as a tool for supporting users during the information seeking process, and especially in exploring a new domain, for refining an information need and formulating good queries. Such a solution is badly needed today, with searching tools widely available to people not trained in how to search. It is not possible to provide a human mediator for every Web searcher, but an automatic mediator could dramatically improve searching effectiveness and, implicitly, the users' satisfaction.

1.5 Road-map to the thesis

Our work relies on a vast amount of research conducted especially in the area of Information Retrieval, but also in Human Computer Interaction, Cognitive Psychology, Computational Linguistics and Statistics. It is impossible to review here all these domains. A cut-down review of the particular areas from which we used results for our work is presented in chapter 2. That chapter is needed in order to understand some of the models and techniques that we have adopted and to understand why we adopted them.

Details of the **mediated access** approach proposed in this thesis are described in chapter 3. We also present the assumptions and the predictions of our model and develop an evaluation framework to test these. The objective of the thesis and its expected contributions are also highlighted.

The design and evaluation of WebCluster, our system that reifies the concept of mediation are presented in chapter 4. The conceptual design of a toolkit and a framework to build mediation systems is also described. Details of the software engineering design are ignored. We concentrate on the flexibility and reusability of the system for building a variety of mediation systems and for supporting a variety of IR experiments.

As clustering is the method proposed for structuring the specialised source collection, chapter 5 is devoted to experiments that test the cluster hypothesis assumption. Moreover, it reviews the original formulation of the cluster hypothesis and proposes the **aspectual cluster hypothesis**, which explains better the results obtained by researchers in cluster-

ing over the years.

Several mediation strategies are proposed and compared in chapter 6, by simulating an ideal user conducting searches on a set of test topics. These experiments estimate the potential of mediation by computing upperbounds of performance with various mediation strategies and comparing them with results obtained through a baseline, un-mediated search. The results are expected to indicate the potential of mediation for improving retrieval performance, and also to provide guidelines as to which search strategies should be incorporated in the operational system.

The final chapter summarizes the research work presented in the thesis, draws conclusions from the WebCluster project and describes future work needed to investigate other aspects of mediated retrieval.

Chapter 2

A Review of Information Retrieval Models and Tools

2.1 Justification

When building an IR system, three essential choices need to be made:

The interaction model - describes the way in which the user interacts with the system, the scenarios (use cases) supported and the level of control over the system during the interaction. It is dependent on the envisaged application and its elaboration should be based on design principles resulted from Human Computer Interaction research.

The retrieval model - describes the indexing process, the representation used for documents and other information holders, the representation of the query, and the matching process between the query and the information holders.

The evaluation model - describes indicators of performance, as well as the methodology for measuring them. Thresholds of minimum quality can be set on performance to validate a new system or, more commonly, the experimental system's performance is compared to that of a baseline system in order to estimate the increase or decrease in quality.

While all three elements are essential in building a new system, or proposing a new approach to information retrieval, they are relatively independent. A novel research con-

tribution could be, for example, a new model underlying indexing or retrieval, such as the *language models* proposed fairly recently [PC98], a new model for evaluation, such as the framework proposed by Borlund and Ingwersen [BI97], or an integrated approach that covers all three areas, such as Jose's work [Jos98]. If a new piece of research contributes to only some areas, it usually uses known results from the complementary areas to build an integral model of the retrieval process and its expected performance.

Our WebCluster project is not an isolated piece of research. It relies on older and on more recent research in Information Retrieval and it is expected to contribute with its results to a better understanding of some areas of IR. The purpose of this chapter is to present the context of our work by reviewing the areas of IR research that have influenced our work, or that we intend to contribute to. Once this context is created, it will be easier to explain, in the next chapter, where the WebCluster project fits within Information Retrieval research.

Space restrictions prevent us from covering all the research areas relevant to our work. We limit ourselves to the fields that have directly influenced our work, or whose results we have used. Therefore, some familiarity with Information Retrieval is expected from the reader. Other readers are kindly directed to more comprehensive reviews, starting with older ones [Sal68, Rij79] and finishing with more recent ones [Kor97, BYRN99, Bel00, KM00].

2.2 Information Retrieval models

2.2.1 Introduction

The purpose of an *information retrieval system* (IRS) is to electronically store documents and to support the users's search for *relevant* documents. Relevance is not an intrinsic property of a document, but is relative to a user's information need, or knowledge gap, or 'anomalous state of knowledge' [BOB82], in a certain context. Relevance is a complex issue, with various aspects to consider [Miz97, Miz96, Par97]. Here a simplified view is taken: a document is relevant for the user if it is 'about' the topic the user is interested in. In order for the system to search for relevant documents it is necessary that:

- the user's internal cognitive state or information need is turned into an external expression or query, based on a *query model.*
- each document is assigned a representation that indicates what the document is *about* and what topics it covers, based on a *document model.*
- a *matching (or similarity) function* can be used to estimate the relevance of a document to the information need, based on the document model and on the query model.

It is important to stress that only an estimation of the relevance of a document to a query is possible. The reason is three-fold. Firstly, articulating a good question, based on an information need, is often the hardest part of answering it [Bel00, p.6]. Techniques for supporting the user in this endeavour, although helpful, are not perfect. Secondly, the words of a document do not unequivocally determine the semantics and the topics of a document. Words may be ambiguous, or rely on context, plus the choice of words for conveying an idea may depend highly on the document's author. Thirdly, there is a variety of matching functions that can take into account a variety of document features such as the distribution of terms, and the size or the style of the document.

The trio *document model - query model - matching function* constitute an *Information Retrieval model.* This section discusses mainly how documents and queries can be indexed in order to obtain representations usable in algorithmic processing. It also looks at some influential models for estimating the relevance of documents for a given user query or set of queries.

Obviously, the document model and the query model need to be compatible in terms of representation (they need to convey in a similar way what a document is about, respectively what the user's interest is) and to afford some measure of similarity between them. Together, the document model and the query model, are commonly known as the *indexing model.* It is the dominant element of the IR model, as it determines the representation of documents and queries and it constrains the range of matching functions that can be used. Therefore, even if the indexing model is only a part of the IR model, the two concepts are often used interchangeably.

2.2.2 Indexing models

An *information retrieval system* stores documents and, based on an information need expressed by a user, retrieves the ones estimated to be relevant. In order for the relevance assessment and the retrieval to be possible, the documents need to be analysed and described, or in other words *indexed*, in order to identify (or at least estimate) their content, meaning, purpose and features.

A lot of recent research has addressed the problem of representing, indexing and retrieving images, sound or other media. However, the review of indexing in this thesis is limited to text documents, the only media considered in the research work described here. Future work may deal with multimedia.

The two approaches to indexing are *manual indexing*, based on human analysis[1], and automatic indexing, based on machine algorithms. Recently, Anderson and Perez-Carballo reviewed and compared the two approaches, looking at various aspects of indexing [APC01a, APC01b]. Their conclusion is that research into comparing the merits of manual vs. automatic indexing has been inconclusive: the advantages and disadvantages of the two tend to level out. The choice of one or the other may need to be made based on the domain (general or specialised), on the specific collection (homogeneous or heterogeneous, small or large), on the specific application, and on practical issues (budget, availability of specialist indexers). Manual indexing is relatively expensive, impractical for very large collections and too rigid to support a large variety of indexing strategies. Automatic indexing is cheap, fast and flexible, but is unable to 'understand' the semantics of documents. Anderson and Perez-Carballo advocate the general use of machine indexing, which is much cheaper and faster, and additional human indexing of *important documents* in order to make them more accessible by identifying themes, relationships, methodological approaches, points of view, prejudices, biases, slants, purposes, values, and qualitative aspects that cannot be easily identified through automatic techniques. These 'important documents' can be identified through use, citation, publisher prediction, reviews and awards, searcher or indexer nomination or advisory board.

[1] "Manual" is a misnomer, used for historical reasons. The human activity of indexing and cataloguing is primarily intellectual, rather than manual.

Although the quality of indexing can potentially influence the effectiveness of Web-Cluster, research into the effect of indexing on mediation is not an objective of this thesis (although it can be followed as future work). What is needed for the purpose of the research described here is an indexing system that is cheap and fast and also sufficiently flexible to be applied to a variety of test collections. Therefore, it is only automatic indexing that is explored further.

Indexing a textual documents consists of identifying 'typical' *keywords* (or *terms*) - words, pieces of words, or phrases - as features that characterise the semantics of the document by conveying the topic(s) covered. The document representative or label is typically represented by keywords that are *representative* for the topics covered in the document, and also have *power of discrimination* in order to distinguish it from other documents in the same corpus or document collection. There are several aspects to consider. The *domain of discourse* or subject area is essential in building document representatives: in a corpus on "artificial intelligence", the phrase "computer science" is too general to be representative, while "machine learning" is a good indication of the sub-domain referred to by a document. The *context* is also essential in considering the vocabulary used for document labels: the audience need to understand the meaning of the keywords. Therefore, "keywords are best viewed as a relation between a document and its prospective readers, sensitive to both characteristics of the users' queries and other documents in the same corpus" [Bel00, p.13].

For now we are not going to investigate the effect of the indexing scheme on the outcome of mediated retrieval. Our indexing algorithm performs the standard stop-word removal, (Porter) stemming and term frequency counting in each document [FBY92]. Future work may also consider phrases or word combinations, or may do a deeper analysis of orthography (hyphenation, capitalization, punctuation, apostrophes, parentheses), for improving the quality of indexing, the estimation of inter-document similarities, and consequently the performance of the system.

Older, boolean systems only recorded the existence or non-existence of terms in doc-

uments, making no difference between contributions of different words to the topic of the document. We would like, however, to be able to distinguish between *content-bearing terms*, which make up the topic, and *context words*, with disambiguation role. Therefore, in our review we are only going to consider more modern indexing models, which assign a *weight* to each term, indicating their significance or contribution to the document representation as well as their power of discrimination, or ability to pick relevant documents out of the many non-relevant documents. Let us investigate some ways in which these weights can be generated.

The vector space model

The *vector space model* [SWY75] represents documents and queries as weighted vectors in an index terms' 'space' whose dimensionality is given by the vocabulary size. The weight of an index term reflects its significance in terms of representativeness and discrimination power. A variety of *weighting schemes* are available for generating the weights. Most of them are based on 3 proven principles [RSJ97]:

1. Terms that occur in only a few documents are often more valuable than ones that appear in many.

2. The more often a term occurs in a document, the more likely it is to be important for that document.

3. A term that appears the same number of time in a short document and in a long one is likely to be more valuable for the former.

These principles generate the *tf-idf-dl* (or sometimes called just *tf-idf*) class of formulae, based on term frequency, inverse document frequency and document length, as described below.

Considering a collection of N documents, the *inverse document frequency* of a term t_i that appears in n documents is:

$$IDF_i = \log N - \log n.$$

The *term frequency* of a term t_i in a document d_j is defined as:

$$TF_{i,j} = the\ number\ of\ occurences\ of\ term\ t_i\ in\ document\ d_j.$$

The *length* of a document d_j is:

$$DL_j = the\ total\ number\ of\ term\ occurrences\ in\ document\ d_j.$$

The document length is usually normalized by the length of the average document:

$$NDL_j = DL_j\ /\ (average\ DL\ for\ all\ documents),$$

or

$$NDL_j = DL_j \cdot \frac{N}{the\ total\ number\ of\ term\ occurrences\ in\ the\ collection}.$$

The generic tf-idf formula is a combination of the measures defined above:

$$W_{i,j} = TF_{i,j} \cdot IDF_i\ /\ NDL_j. \tag{2.1}$$

In practice, a more complicated formula is used, usually with parameters (or so called tuning constants) that control the effect of the three measures according to the particulars of a collection. For example, Sparck-Jones and Robertson [RSJ97] use the version:

$$W_{i,j} = \frac{TF_{i,j} \cdot IDF_i \cdot (K1+1)}{K1 \cdot ((1-b) + b \cdot NDL_j) + TF_{i,j}}. \tag{2.2}$$

$K1$ controls the extent of the influence of term frequency. The optimal value depends on the length and heterogeneity of documents in the collection and can be set for a collection after systematic trials. The constant b, which ranges between 0 and 1, controls the effect of the document length: it can be set to 1 on the assumption that documents are long because they are repetitive, and to 0 on the assumption that they are multi-topic. High intermediate values indicate a verbose style of document.

Another widely used version of the *tf-idf* is used in the Inquery system [CCH92, PC98]:

$$W_{i,j} = \frac{TF_{i,j}}{TF_{i,j} + 0.5 + 1.5\,NDL_j} \cdot \frac{\log \frac{N+0.5}{n}}{\log(N+1)}. \tag{2.3}$$

We have used this Inquery version of *tf-idf* in some of our experiments because it has proved effective in TREC experiments, but also because it has no tuning parameters that would introduce extra variables in the experiments. We will refer to it as "the TfIdf formula" in the description of the experiments.

Note that a vectorial representation of documents can also be used with boolean values, indicating the presence or absence of the corresponding vocabulary term in each document. This is typical for boolean systems, but can also be seen a particular or simplified case of weighted representation, supporting a ranked retrieval system.

The estimation of the degree of relevance of a document to a query is computed by placing the query, also modelled as a vector, in the same multi-dimensional space and computing some similarity or, alternatively, distance measure between it and every document. Since the query is typically short compared to the document, an angular measure such as Cosine, which prefers the topical content of the vectors, rather than the size, is usually preferred. Relevance feedback is also supported by adding to the initial query the vectors of the documents marked relevant, possibly multiplied by a scalar parameter that controls the influence of RF.

The vector space model is popular in Information Retrieval: it is supported by a strong theoretical model and, despite its simplicity, it has proved highly successful in experiments and in operational systems[SM83, BYRN99]. It is often the model of choice used in document clustering, as it offers an intuitive way to compute similarity between documents: the inter-document similarity can be modelled by angular or Euclidian distances between vectors.

Probabilistic models

The *probabilistic models* rank documents in a collection according to their estimated probability of relevance to the query [Rob77, CLRC98]. The estimations are based on the frequency of query terms in each document. They rely on user *relevance judgements*, employed as training data, to provide typical term frequency distributions in relevant, and respectively non-relevant documents.

For a term t_i, if

$$r = number\ of\ known\ relevant\ documents\ term\ t_i\ occurs\ in,$$

$$R = number\ of\ known\ relevant\ documents\ for\ the\ current\ topic,$$

and with N and n defined as in the previous subsection, then the *relevance weight* [RSJ97] can be computed as

$$RW_i = \log \frac{(r+0.5)(N-n-R+r+0.5)}{(n-r+0.5)(R-r+0.5)}. \tag{2.4}$$

The corrections represented by the 0.5s are due to the polythetic relationship between documents and words: there are no words that appear only in relevant documents and words that appear only in non-relevant documents. A word could appear in relevant documents just because it appears in many documents. Also, the fact that a word is not in any of a few relevant documents does not imply that it will never be.

The relevance weight (2.4) can replace the inverted document frequency in the weight formula (equation 2.2) in a second or subsequent retrieval iteration (or even as a first iteration if R and r are set to 0) to give the *iterative weight*:

$$IW_{i,j} = \frac{TF_{i,j} \cdot RW_i \cdot (K1+1)}{K1 \cdot ((1-b)+b \cdot NDL_j) + TF_{i,j}}. \tag{2.5}$$

Note that in principle RW_i can replace IDF_i in the generic formula of *tf-idf* (2.1) or in its Inquery form (2.3). However, as the relevance weight was proposed by Sparck-Jones and Robertson, who pioneered work in probabilistic models, we also used their version of *tf-idf*.

The overall score for the document d_j when matching against a query is simply the sum of weights of the query terms present in the document. For long queries, the *query frequency* of the terms,

$$QF_i = the\ number\ of\ occurrences\ of\ term\ t_i\ in\ the\ query,$$

is multiplied by the document term weights before the addition. Again, despite the difference in the conceptual model, the practical formula is very similar to the vector model's cosine measure of similarity between the document and the query.

When no relevance information is available, initial document scores are computed based on *tf-idf*-type weights applied to the query terms[CH79]. Following user feedback, the weights of the query terms are adjusted in an interactive cycle based on the equation 2.4. Conceptually, the query is viewed as an imperfect approximation of the user's information need and its adjustments, based on user feedback, are expected to improve the approximation and consequently the quality of the retrieved set.

Probabilistic models have become the dominant model used in experimental systems, most participants in the TREC benchmarking exercise using them in some form. In interactive systems it is the users that provide relevant feedback, while for ad-hoc (non-interactive) tasks training data in the form of relevance judgements are used for tuning the parameters in the formulae.

Language models (LM)

Some researchers, such as Ponte and Croft [PC98] have proposed the idea of ignoring relevance judgements in generating the retrieval set and trying instead to concentrate on satisfying the query. Their approach is to describe the 'aboutness' of documents and of queries based on statistical *language models*, i.e. on distributions of term frequencies. The output of the IR system is a ranked list of documents, the score of each document being the probability that the query was derived from that document. In other words, considering the document a population of words, the issue is to estimate the probability that the query, seen as a sample of keywords, is a representative sample for the document.

The challenge is to make the right corrections for this model, corrections needed due to the vocabulary mismatch. Firstly, a document is not a complete population of terms, the author choosing familiar words, when other words would have expressed the same 'topicality' or 'aboutness' of the document. The fact that a word highly specific for a topic does not appear in a document is not necessarily an indication that the document

does not cover the topic; a less specific synonym may have been used instead. Secondly, the query is imperfect, especially when the user is not a domain expert - the user has chosen certain words, when others may have been equivalent or better. The fact that a word is not in the query is not necessarily an indication that the word is not a good contributor to describing the user's information need.

The language model (LM) approach is relatively new and unproven. Initial claims of significant increase in retrieval effectiveness [PC98] were not confirmed by further experiments [SC99]. Also, there is no concord yet with regards to the smoothing techniques to be used for model corrections. Even the need for smoothing techniques for correcting the query, viewed as an imperfect formulation of an information need, may be seen as a flaw in the model: the correction attempts to make the query a better formulation for the information need, while the theoretical model separates the query from the information need and the notion of relevance. Moreover, LM's are promoted as a 'clean' model, in which statistical information is integrated in the model, rather than used heuristically, like in probabilistic models. However, heuristics are used in smoothing the model.

The new approach has certainly produced an important impact in Information Retrieval, as proved by a recent workshop that brought together leading researchers in the field [CCL01]. On the one hand, the LM has been shown to give very good results in areas such as ad-hoc retrieval, personalization, summarization, topic discovery and compression. On the other hand, it has been criticised for flaws in its conceptual model, for subtle contradictions between the theoretical model and its application, and for its (in-)ability to deal with relevance feedback. Current work on language models attempts to address weaknesses in the LM model or to extend it, especially in terms of integrating relevance feedback [Pon00, XC00, Hie01, RH01, LC01].

A major contribution of the new approach is the fact that it connects Information Retrieval to Information Theory and Speech Recognition research and it offers IR researchers tools and results from these fields. One such tool is the *relative entropy* or *Kullback-Liebler divergence*,

$$KL = p \log \frac{p}{q}, \tag{2.6}$$

which measures how well a probability distribution predicts another or, on the contrary, how different the two probability distributions, p and q, are [MS99]. Xu and Croft [XC99] proposed two ways in which this measure can be used. For retrieval purposes, the KL divergence can be used to measure how well a topic model for topic T predicts a query Q:

$$KL(Q,T) = \sum_{f(Q,w_i)\neq 0} \frac{f(Q,w_i)}{|Q|} \log \frac{f(Q,w_i)/|Q|}{p_i}, \tag{2.7}$$

where $f(Q,w_i)$ is the number of occurrences of term w_i in Q, $|Q|$ is the length of Q in words, and p_i is the probability that the term w_i describes the topic T. If D is a set of documents describing topic T, the p_i is estimated as:

$$p_i = \frac{f(D,w_i) + 0.01}{|D| + 0.01 \cdot n},$$

where $f(D,w_i)$ is the number of occurrences of w_i in D, $|D|$ is the size of D in words and n is the vocabulary size. The correction (the authors do not justify the value 0.01) is made for practical purposes, for query terms that do not appear in any document, and is justified by the fact that D is just a sample, and not the set of all possible documents, describing the topic T.

Xu and Croft applied formula 2.7 to cluster-based retrieval, i.e. the retrieval of clusters that have a high probability to generate a given query. If D is reduced to a single document, the formula can be applied to ranked retrieval.

Another use of the Kullback-Liebler formula that Xu and Croft propose is for clustering[2]: in the iterations of a K-means clustering method, the dissimilarity between a document d and a cluster c can be computed as:

$$KL(d,c) = \sum_{f(d,w_i)\neq 0} \frac{f(d,w_i)}{|d|} \log \frac{f(d,w_i)/d}{(f(c,w_i) + f(d,w_i))/(|c| + |d|)},$$

[2]Clustering is reviewed in the next section.

where $f(d, w_i)$ is the frequency of the term w_i in document d, $f(c, w_i)$ is the frequency of the term w_i in cluster c, $|d|$ is size of d and $|c|$ is size of c.

2.2.3 Comments

Although the conceptual model, the assumptions, the theory and the justification for the formulae of all the models described look very different, there is strong commonality between them. While not attempting to achieve real understanding of semantic meaning in documents, they all look at statistical information (mainly term frequencies) in order to capture the topic(s) of the documents and to estimate the relevance of each document for a query or an information need. They all return a ranked list of documents, with the top documents expected to be more useful for the user than the bottom ones: in vector space models the top-ranking documents are those 'close' to the query; in probabilistic model those highly probable to be relevant to the query; and in statistical language models those that are highly probable to generate the query. All the models are based on the distribution of term frequencies in the documents and in estimating some commonality between documents and the query, based on various estimation and data smoothing techniques.

The underlying commonality allows for some careful mixing of the models in an IR system, based on a common representation of data. For example, document similarity in view of clustering can be based on the vector model or computed by comparing language models of the documents. In the same system ranked retrieval can be based on the probabilistic model, or on a language model, or on a vector space model. The possibility of combining models in a pragmatic fashion is useful because the models differ in predictive power and in the way the estimations can be improved, so the ability to use the model that is the most appropriate and advantageous in a certain situation is important.

Before closing this section, we need to highlight a subtlety that may emerge when mixing models. The weighting of document terms is usually seen as part of the indexing process in most systems based on the vector space model. The weights are derived from term frequencies and the system builds and stores document representations based on these weights. On the other hand, most systems based on the probabilistic model see the weighting of terms as part of the retrieval process. This is natural, as the weights change

based on iterative relevance judgements.

If we want to mix and match models, we need to reconcile these distinct views. A solution, which we used in WebCluster, is to store 'permanent' document representations based on term frequencies. If a probabilistic model is employed, terms weights can be derived from frequencies. If a vector space model is used, then weights can be computed and 'working' document representations can be derived during retrieval even if, conceptually, this step can be viewed as part of indexing. This approach also allows flexibility in applying a certain weighting scheme at retrieval time, rather than fixing it at indexing time.

2.3 Document clustering

2.3.1 Structure and Information Retrieval

The phenomenal increase in the quantity of information made available electronically does offer, as some enthusiasts of the information age put it, "power at our fingertips". However, faced with an information need, a user can easily become overwhelmed by the sheer quantity of information and unable to distinguish between what is relevant and useful for the task at hand, and what is not. Undoubtedly the most successful approach to organising this mass of information in order to help the user make sense of it is to provide *structure.*

Encyclopaedia Britannica provides an example of successful organization of information, which affords both exploration in view of learning, and search for references to information. Its *Propaedia* contains a hierarchically structured outline of knowledge and a guide to the contents and use of the Encyclopaedia. It gives an indication of which subjects are covered and references to more in-depth information. *Micropaedia* offers a list of articles covering topics in human knowledge, sorted alphabetically based on the keyword that represents the topic. Subjects that require more in-depth treatment are covered by the *Macropaedia.* Finally, the *Index* supports keyword-based searching for topics.

Such an organization of information, improved and proven to be effective over the

centuries, could be emulated in IR systems. A *hierarchical* organization of the topics can support either learning, for the user unfamiliar with a domain, or better navigation, for the user knowledgeable of the semantic structure of the domain. To complete the picture, *referential* links offer connections to more coverage of the topics of interest, and search engines play the role of the Index. While referential links have proven their utility in hypermedia applications and particularly on the Web, we concentrate here on methods for hierarchically organizing information in support for learning or semantic navigation.

Traditionally, the structure of the information space was used for improving the efficiency or effectiveness of the retrieval algorithms, or for automatic query expansion. The more recent advent of interactive information seeking environments has increased the importance of structure and of classification methods, and research in Human Computer Interaction and Information Visualization has brought an important contribution by offering visualization and navigation tools appropriate for exploration.

Both manual and automatic methods have been proposed and investigated for classifying document collections in view of supporting information retrieval [Hea99a]. Manual classification is typically optimised for a specialised domain, capturing the domain experts' consensus or compromise, so it is likely to better reveal the semantic structure of the domain, its main topics and their subtopics. On the other hand, clustering has the advantage of being fully automatic and therefore faster and cheaper. It is also domain independent, data driven (it relies on the actual content of the documents rather than on expert knowledge about the domain) and usually successful at identifying meaningful themes in relatively heterogeneous collections.

Observations of searchers' behaviour in libraries show without a doubt the importance of *structure* in learning, in exploring a domain, or in conducting a search. We therefore envisage the use of structure for guiding the user's exploration as part of the mediation paradigm. If the specialised collection used for mediation has a taxonomy associated with it or has been classified manually, then that structure can be used to support navigation. However, not every problem domain has a taxonomy or a classified collection associated with it. Such collections can be clustered and the obtained structure can be used to guide

a search.

The mediated access approach, proposed and promoted in this thesis, relies on structure to support exploration in an interactive setting, no matter how the structure is obtained. In an operational system based on mediation any method that does a good job of revealing the topical structure of a document collection is appropriate. Our research will concentrate, however, on the use of document clustering and this is for two reasons:

1. The use of clustering means that the full mediation process (both the preparatory stage of indexing and structuring of the specialised collections, and of generating document and cluster representatives, as well as the operational, interactive stage) are fully automatic and independent of domain.

2. We have a particular interest in clustering and wish to contribute to research in this direction.

We will therefore explore document clustering as a tool for structuring a document collection in view of mediation.

2.3.2 Clustering and the cluster hypothesis

Cluster analysis, or *clustering*, is a technique for multivariate analysis that assigns items to automatically created groups based on the calculation of the degree of association or similarity between items, and groups. It has a variety of applications in Information Retrieval, such as grouping together terms (term clustering [SJN68]) based on their collocation, in order to build thesauri, or grouping information sources, user profile and other such 'objects' based on some measure of *similarity*. However, its most common application, and the only one we are going to consider in this review, is *document clustering*.

The initial aim of introducing document clustering in Information Retrieval was to increase the efficiency of retrieval: after an initial overhead, consisting in grouping of documents in *clusters*, based on their reciprocal similarity, the search would have to look for the best clusters of documents that matches a query and not for individual documents [Sal68, Wor69].

It was Jardine and van Rijsbergen who first suggested that the associations between documents convey information about the relevance of documents to requests, formulated the **cluster hypothesis** - "Closely associated documents tend to belong to the same clusters and to be relevant to the same request" - and experimentally showed that *cluster-based retrieval* may yield better results than best-match retrieval [JR71].

In this section we will examine the details of clustering, the outcome of the expectation it raised, and the current use of document clustering. The readers interested in details of the early work on clustering are directed to Willett's review [Wil88].

2.3.3 Attributes

The objects in a collection to be clustered need to be described by *attributes* or *features*, so that the similarity between them can be estimated. In the case of document clustering (which can be applied to full documents or to document surrogates such as abstracts or summaries) it is usually the keywords that constitute the attributes [Kwo75, HZ80]. Therefore, clustering is actually applied to document representatives, obtained through indexing, as explained in sub-section 2.2.2.

Other features such as citations or author(s) co-occurrence can also be used as a measure of association [BP74, SS85]. Multimedia documents can be represented by a variety of attributes, specific to various media. Of course, multiple sets of attributes (such as annotations and spatial features, in a photographic collection) could be simultaneously used and their contributions to computing document similarities could be combined by, for example, using the Dempster-Shafer theory of evidence combination [JIH96, Jos98]. In principle, clustering works in the same manner as for text documents, as long as document representatives can be generated and some measure of similarity can be computed between them.

Our research only deals with text documents indexed automatically. In the rest of the thesis, "document" refers to "text document" or, more precisely, to "document representative".

2.3.4 Weighting schemes

As discussed in the review of indexing models, weighting schemes have been used to capture the contribution of index terms to each document in a corpus. Schemes such as *tf-idf* have been conclusively shown to improve retrieval effectiveness.

The effect of weighting schemes on document clustering is less clear. Intuitively, they affect the representation of documents and implicitly the similarity between documents, so 'good' weighting schemes are expected to more accurately identify similar documents, so that they can be grouped together by clustering algorithms.

Among the few experimental results that have been reported in this area, Willett's conclusions indicate that weighting schemes do not lead to a consistent improvement in performance over the use of unweighted terms [Wil83]. However, these results and the consequence that sophisticated weighting formulae are not worth the effort are limited by the choice of similarity measures tested, by the test collections used in the experiment, and by the measure of performance employed. These results are challenged, for example, by Dubin's hypothesis that the capacity of clustering to group together topical documents depends on the power of discrimination of the terms that make up the document representatives. Although the results obtained by Dubin in the context of using clustering for building browsing spaces are somewhat inconclusive [Dub96], it may be worth examining the effect of weighting schemes when clustering is used in a new context, namely mediated access.

Apart from arguing in favour of more experiments on the effect of *internal* weighting schemes (based on term frequencies in the documents), we also suggest that an *external* one may be used to beneficial ends. We accept the argument that some collections may well have a natural structure, in which the position of each document is quite clear, in which case clustering could reveal the 'inherent' structure, rather than impose a convenient structure. However, for a collection covering a variety of complex, multi-faceted topics, with the documents potentially covering several topics or aspects, each user may be interested in a certain view, or topical projection of the collection. Therefore, an external scheme, for example extracted from *frequently asked questions*, may be able to bias

the structure in the direction wanted by the user by influencing the major and the minor axes of the clustering.

2.3.5 Similarity/dissimilarity coefficients

Similarity/dissimilarity coefficients are functions that associate a real value to a pair of items in the collection, based on the attributes that describe them, indicating the degree of similarity/dissimilarity or 'likeness'/'unlikeness' between them. Ellis et al. [EFHW93] did a good job of reviewing these coefficients, so this section will only cover the minimum necessary for supporting our work.

Intuitively one would expect that the more terms two documents have in common, the higher the similarity between them. In most practical situations it makes sense to also take into account the weight of the terms, so that highly specific terms contribute more to the similarity calculation than common terms. Therefore, if $X = (x_1, \ldots, x_v)$ and $Y = (y_1, \ldots, y_v)$ are vectors representing the two documents, where v is the vocabulary size, the *dot product* gives a good indication of the inter-document similarity:

$$X \cdot Y = \sum_{i=1}^{v} (x_i \cdot y_i). \tag{2.8}$$

The dot product is usually *normalised*, so that the similarity formula yields values between 0 (indicating no similarity) and 1 (indicating complete similarity). Two widely used normalised coefficients are:

$$Cosine(X, Y) = \frac{\sum(x_i \cdot y_i)}{\sqrt{\sum(x_i)^2 \cdot \sum(y_i)^2}} \tag{2.9}$$

and:

$$Dice(X, Y) = \frac{2 \cdot \sum(x_i \cdot y_i)}{\sum(x_i)^2 + \sum(y_i)^2}, \tag{2.10}$$

but many others are described in IR textbooks.

Cosine is particularly important for clustering due to its intuitive interpretation (the cosine of the angle between the two vectors in the document space) as well as its proven effectiveness. The direction of a vector can be viewed as the topic of the document it represents, so a high score with the Cosine similarity measure indicates documents that

are expected to cover the same topic.

It is possible, and sometimes easier or more intuitive, to use *distance coefficients* indicating dissimilarity, instead of similarity coefficients (with which they are in a complementarity relationship) [Rij79]. Their advantage is a simple geometric interpretation, convenient for graphical representations, so they are popular with visualization tools. The disadvantage is that they can give an inaccurate representation of reality and lead to two documents being regarded as highly similar even if they have no common terms[3]. Although favoured by taxonomists, due to their relative ease to visualize, they are not widely used in document clustering with the exception of the Euclidean distance,

$$Euclid(X,Y) = \sqrt{\sum(x_i - y_i)^2}, \tag{2.11}$$

used, for example, in Ward's method, as discussed in sub-section 2.3.6.

2.3.6 Clustering methods

There is a variety of classifications for clustering methods [SS73, CO90]. From the point of view of supporting exploration, and implicitly mediated access, it is sufficient to distinguish between hierarchic and non-hierarchic clustering methods. These two classes of clustering methods are discussed in more detail in the rest of this section.

Non-hierarchic clustering methods (NHCM)

These divide a collection of N documents into M clusters, and use heuristics in assigning documents to clusters in order to give very good computation efficiency. Used recursively, they can also be used for generating hierarchic structures.

In most cases the classifications obtained depend on the order in which the documents are processed and on the heuristic parameters used. In cluster-based retrieval experiments most of them showed a drop in retrieval effectiveness compared to ranked retrieval [JR71, Wil88], so they were deemed unsuitable for use in Information Retrieval.

[3]For example, using the Euclidean distance the distance between the documents (00010) and (00001), that do not have common terms, is the same as the distance between the documents (11110) and (11101), that have 3 terms in common.

More recently however, the power of visualization and of browsing and navigation tools has been recognized, particularly in the context of World Wide Web retrieval. Due to their efficiency in terms of speed and memory, non-hierarchic clustering methods have regained attention, due to their capacity to organize search results fast [CKPW92]. Although considered less well-performing than hierarchic methods with regards to effectiveness, or quality of clustering, partitioning methods have been shown to be adequate in retrieval tools. Such tools can support the user's exploration of the search output by conveying the structure of the retrieved set of documents [PSHD96] and can significantly improve retrieval performance [HP96].

Algorithm 1 Single-pass partitioning.

```
The first object becomes the cluster representative of the first cluster.
for all new object do
  Match object against all existing cluster representatives.
  if all matching values are under a certain threshold then
    The object becomes the cluster representative of a new cluster.
  else
    Assign object to best matching cluster (or to more if overlap is allowed).
    Recompute the representative of that cluster(s).
  end if
end for
```

The generic procedure for *single-pass* clustering [Rij79, p.52] is captured in Algorithm 1. There is no direct control over the number of clusters and the clusters' size, although the condition of the matching function can be modified for an indirect control.

Algorithm 2 Iterative partitioning.

```
Select M objects as seeds for the M clusters.
for all new object do
  Assign object to a cluster according to some matching function.
  Recalculate the seeds.
end for
Iteratively shuffle the objects between clusters, trying to maximize some objective function.
```

The generic procedure for *iterative* clustering is presented in Algorithm 2. In general, the number of clusters is established a priori, heuristically. There are different implementations that vary in the way the initial seeds are established, in the way the seeds are

recomputed (at the end of each iteration or after an object is assigned to a cluster, for example), in the way the number of clusters is established, in the way objects are shuffled, and so on.

An exception among these mostly heuristic clustering methods are the methods proposed by Can, based on the *cover-coefficient* concept [CO83, CO85, CO90, Can93], which estimate the number of clusters and their sizes based on the attributes of the objects in the collection, and which stand out through their theoretical soundness and their advantages:

1. The number of clusters and their sizes can be estimated from the attributes of the objects in the collection.

2. They distribute the objects uniformly among clusters - they do not cause a few 'fat' clusters and a lot of singleton (or very small) ones.

3. They are independent of the order of the documents.

4. Their complexity (and, therefore, efficiency) is better than most other clustering algorithms.

5. Their retrieval effectiveness is comparable to the one of complete link algorithms.

Hierarchic clustering methods (HCM)

HCMs create tree-like classifications in which clusters of highly similar documents are nested within larger clusters of less similar documents. The single cluster containing the entire collection is represented by the root of the tree while the individual documents reside in the leaves; the other nodes correspond to clusters at different levels of similarity. Hierarchic structures are a familiar concept, found in real life taxonomies and employed by domain experts when manually classifying collections. Therefore, we favour hierarchic clustering methods over partitioning methods as techniques for structuring collections in view of supporting exploration.

These methods can be classified as:

Divisive - the single initial cluster is divided into smaller and smaller clusters of documents, by finding dissimilarities between documents within clusters. They usually

create *monothetic* classifications, in which all documents in a cluster must contain certain terms in order to belong to it.

Agglomerative - the cluster structure is built by successive fusions of clusters, starting with each document in a singleton cluster. They usually create *polythetic* classifications, where documents in a cluster have terms in common, but there are no specific terms required for cluster membership.

The hierarchic agglomerative clustering methods (HACM) are the most popular clustering methods for information retrieval because, after an initial overhead represented by the computation of the inter-document similarities and the building of the cluster structure, the retrieval is efficient and effective.

Algorithm 3 Hierarchic agglomerative clustering.

```
Each item to be clustered constitutes a singleton cluster.
Compute similarities between clusters.
while there is more than one cluster do
    Merge the most similar clusters.
    Recompute similarities between clusters.
end while
```

The generic procedure used by such methods is captured by Algorithm 3 [Voo86]. It is apparent that such a method is defined by the choice of the *(inter-) cluster similarity*, i.e. the measure used to calculate the similarity between two clusters. It can be a formula similar to the document similarity measure, if the similarity is calculated between cluster representatives (see section 2.3.10), or it can be derived from the similarities between constituent documents.

Divisive methods have received less attention [SKK00] and are less commonly used, so we will concentrate on the agglomerative methods. The most widely used HACMs are described below.

Single-link clustering method

In this method the inter-cluster similarity is defined as the similarity between the most similar pair of documents, one from each cluster. There are several theoretical advantages [JR71] over other methods:

- The clustering obtained only depends on the rank-ordering of similarity values, not on their absolute values.
- It is stable under small errors in similarity values.
- It is stable under update: the cluster hierarchy is unlikely to change drastically when further objects are incorporated.
- The order of input is not significant. A given set of data should define exactly one hierarchy.

but also some disadvantages:

- It tends to form long, loosely bound clusters with little internal cohesion (phenomenon known as *chaining*), with documents in the same cluster not necessarily more similar to each other than to documents not in the cluster.
- It produces a high number of *aberrant* documents, documents not similar to other documents, isolated at highest levels in the hierarchy [JR71].

Complete link clustering method

The inter-cluster similarity is defined as the similarity of the least similar pair of documents, one from each cluster. It tends to create small, tightly bound clusters containing highly relevant documents. Certain algorithms that implement it, combined with certain search strategies give the best levels of effectiveness obtained in cluster-based retrieval experiments, although they are also the most demanding of computational resources [Wil88].

Group average clustering method

This method is based on the mean of similarities between all pairs of documents, one from each cluster. It is an 'average' method with regards to efficiency and effectiveness of retrieval, but has very good stability and recovery characteristics (see section 2.3.11).

Ward's clustering method

Those clusters are fused that result in the least increase in the sum of distances from each document to the centroid of its cluster. It tends to create spherical clusters which may not

accurately reflect the 'true', or natural, shape of the clusters present in the data set. It is close to the group average method with regards to efficiency and effectiveness of retrieval as well as to the stability and recovery characteristics. Some constraints of this method are that it has to use a Euclidean distance measure and that the cluster centroid has to be re-computed, during the clustering process, whenever a cluster is modified.

2.3.7 Clustering algorithms

General considerations

The difference between clustering methods and clustering algorithms is rather subtle and sometimes indistinguishable. The general acceptance is that methods describe algorithmic steps in a generic fashion, without reference to implementation details such as the data structure employed, or the use of memory and computer storage. Each method can be implemented by a variety of algorithms, which differ in terms of:

- the data structure - a specialised data structure can improve access speed, but increase the complexity of the code and decrease the flexibility and adaptability of the algorithm to different collections or other conditions.

- the use of memory - lower computational complexity and higher speed can be obtained if the full data structure is stored in memory and if the memory is also used for intermediary results. The drawback is the need for computers with large memory or, alternatively, a severe limitation in the size of the collection to be clustered.

- the use of storage - the use of storage is usually balanced against the use of memory. However, for very large collections it is infeasible to store full inverted files or other structures in memory, so fast access to data on disk is essential. The use of appropriate data structures, of caching and of compression can make a substantial difference [MZ97].

Even if they differ in time and storage complexity, algorithms that implement the same clustering method are expected to generate the same result to a given input. This is not always the case, as the algorithms may use different thresholds or tuning parameters in order to increase efficiency or effectiveness, which may (hopefully only slightly) alter the results.

The goal of the research work described in this thesis is to propose the novel approach of mediated access and to build a proof of concept based on clustering. Finding the best-performing methods and algorithms and tuning parameters is not a major objective. Therefore, we will limit ourselves to implementing and evaluating the 'classical' methods reviewed so far in the context of mediation. We are particularly interested in hierarchic clustering methods, which build a structure more appropriate for exploration in an interactive setting. The investigation of other methods may follow in the future.

In the future we may want to investigate other methods, especially ones which attempt to identify topics and to cluster documents according to *topics* or *concepts* rather than word frequency similarity[4]. With the advent of the Web and multimedia, we should also be looking at methods that address distributed collections, large collections, and media other than text.

Combinations of algorithms

In order to combine specific advantages that they present, various algorithms can be combined, especially when no single method can be applied with satisfactory results. For mediated retrieval such a case is the structuring of very large specialised collections in view of exploration. Hierarchic clustering methods are infeasible, due to their complexity. On the other hand, partitioning methods would produce either too many or too large clusters to make exploration feasible. Two solutions, originally proposed by Croft [Cro77] and respectively Jardine and van Rijsbergen [JR71], can be adapted for this situation:

- Applying a fast single-pass non-hierarchic clustering algorithm in order to partition the collection into a number of big clusters on which a hierarchic clustering method can subsequently be applied. A disadvantage is that single-pass clustering methods are heuristic and depend on the input order of documents, so they may degrade the quality of clustering.

- Doing a *core clustering*, on a sample subset of documents, followed by the assignment of the other documents to the resulted clusters, using a document-cluster matching function and a downward search strategy. Some problems are finding a *representative*

[4]This may be a matter of changing the indexing algorithm and the inter-document similarity function, rather than the actual clustering methods.

and large enough sample for the core clustering and the fact that adding documents by using a heuristic search strategy may degrade the quality of clustering.

2.3.8 Efficiency issues

Conceptually, the similarity values between each pair of documents in a collection form a *similarity matrix*. Of course, it is a symmetric matrix and the values on the diagonal correspond to maximum similarity (each document is perfectly similar to itself).

Various HACM algorithms deal differently with the similarity matrix [Wil80]. Some calculate and store the similarity matrix (either the upper or the lower part of the symmetric matrix) prior to clustering, while others save space (but increase computational complexity) by calculating a similarity value whenever it is needed. Some compute the whole similarity matrix, while others only compute values estimated to be significant, which can improve efficiency in the case of sparse similarity matrices (when each document has a relatively low number of similar documents).

Inverted files [Rij79], usually built during the indexing of the documents, offer clues as to which similarity values do not need to be calculated: the similarity between two documents that have no terms in common is obviously zero. Algorithm 4 is such an algorithm that uses an inverted file to avoid computing the zero values in the similarity matrix [Ras92].

Algorithm 4 Matrix reduction - Rasmussen version (notation changed).

```
for all doc1 in the collection do
  for all term in doc1 do
    retrieveInvertedList(term)
    for all doc2 in invertedList do
      increment counter[doc2]
    end for
  end for
  for all doc2 in the collection do
    if counter[doc2] ≠ 0 then
      calcSimilarity(doc1, doc2)
    end if
  end for
end for
```

This is just a conceptual algorithm, the actual implementation allowing for several optimisations. It is sufficient to compute half the matrix, for example only the values $sim_{i,j}$, with $i < j$. Another improvement could be having the traversal of the inverted file (term by term) drive the main loop of the algorithm.

The idea of using the inverted file to calculate a sparse similarity matrix, proposed by Croft [Cro77], works well in terms of efficiency when short document descriptions are used. In the case of indexing exhaustivity, when documents are represented by a large number of terms, a large number of non-zero-valued coefficients are repeatedly calculated, as pairs of documents are on the posting list associated to each index term they share, with a substantial increase in running time [HW80]. Willett improved this algorithm, avoiding the calculation of redundant similarity values [Wil81].

Croft has also shown that by ignoring the longest posting lists in the inverted file[5], containing the most common index terms, with low discrimination value, the number of similarity values calculated can be reduced significantly [Cro77]. Of course, such action introduces an error, which was experimentally shown not to affect the outcome of clustering. A similar action would be to alter the algorithm above in the sense of setting the zero value to the similarity between each pair of documents whose number of common document is under a certain threshold.

A further reduction of the number of similarity values to be calculated is by considering only the k nearest neighbours of each document and ignoring the other similarities. The so called *nearest neighbours* clustering, advocated by Croft [Cro78], has been largely employed in document clustering - Willett [Wil84] and Smeaton [SBCQ98] have shown that this approach brings a great improvement in efficiency, without a loss in effectiveness. It must be said that most of these experiments ignored the effect of the approximations on the similarity matrix and on the structure generated through clustering. Only the effect on the effectiveness of cluster-based retrieval was measured.

Willett went further [Wil96], advocating the use of *nearest-neighbour* clustering (in

[5]Croft ignored the first two in his experiments.

which each cluster contains just 2 documents, they being each other's nearest neighbour), with good efficiency compared to hierarchical agglomerative clustering methods and effectiveness comparable with and complementary to, conventional best-match searching.

Another issues is the inter-cluster similarity matrix, containing the similarity between each pair of clusters. Conceptually, this matrix is different from the inter-document similarity matrix. However, for agglomerative clustering methods, one matricial data structure is sufficient. Initially it contains the computed similarities between the documents, which are exactly the similarities between the singleton clusters, at the beginning of the clustering process. When two clusters, K and L, are fused by the clustering algorithm, the entries in the similarity matrix corresponding to the new cluster KL can overwrite those corresponding to K and L, so no additional storage is required during clustering. Secondly, the entries for KL can be calculated from the entries for K and L, based on the Lance-Williams formulas [Ras92], without the recourse to the original document similarity matrix.

2.3.9 Search strategies

One application of clustering is cluster-based retrieval: once the cluster hierarchy has been built, a search for the cluster or clusters that best match a query can be done. There are several strategies:

Top-down - the search enters the tree via the root and moves down the tree following the path of maximum similarity. The search can be described as a series of *correlations* and *expansions* [MM72]. The query is compared to all nodes on the current level of the hierarchy (correlation) and one or more of the nodes are chosen based on some *decision criterion* to be replaced by their children (expansion). The search is stopped by some *retrieval criterion* or *halt criterion*, for example when the cluster size drops under a certain value, or when the query-cluster similarity begins to decrease. When the search is terminated, the current cluster is retrieved in its entirety. Depending on the purpose of the retrieval, the documents in the retrieved cluster may be ranked for the final output.

The search can be *narrow*, when a single subcluster is chosen at each level in order

to continue the search, or *broad*, when more than one choice can be made. The narrow top-down search is *precision-oriented*, while the broad one is *recall-oriented*. Both have a complexity of $O(\log n)$, unless a form of backtracking is used, which would increase the complexity.

There are a number of variations of this general procedure. *Forward* search strategies allow only one opportunity to expand nodes on each hierarchy level, while some systems allow *backtracking*; that is, restarting the search on an upper level when an earlier expansion turns out poorly. A *plunging* strategy is a combination of narrow forward search combined with backtracking. Generally, strategies that involve backtracking are worthwhile only if the decision criterion on each level is flexible; if a fixed number of nodes are expanded, a forward search strategy is just as good.

Particularly for large and heterogeneous collections, the topics of the clusters at the top of the hierarchy may be rather vague, so the first few choices in a top-down search may be almost arbitrary. A thresholding procedure could be applied, ignoring the clusters of size above a certain value. Alternatively, the top few levels of the hierarchic structure can be ignored from the search.

Bottom-up - the search moves from a document or a bottom-level cluster towards the root of the tree, the stopping condition being given, for example, by the size of the current cluster. The crucial problem is the starting point - it can be found by a conventional best-match search or by a bottom-level scan of the tree.

For improved precision, a bottom-up search can be followed by a top-down one.

Bottom-level scan - the bottom-level clusters[6] are scanned and the most relevant to the query is retrieved. In Croft's experiments [Cro80], and in the ones conducted with van Rijsbergen [RC75] this strategy gave the best effectiveness, results confirmed by El-Hamdouchi and Willett [EHW89].

If the size of the retrieved cluster is smaller than desired, the search can be continued with a bottom-up search, as proposed by van Rijsbergen and Croft, or more bottom-level clusters are retrieved, as advocated by Griffiths et al. [GLW86]. For the latter case, a further modification, which according to Voorhees gives better results

[6]The bottom-level cluster of a document is the smallest non-singleton cluster that contains that document.

[Voo85b], is to individually rank the documents in these clusters against the query.

Global - In an experimental setting, this approach can be used for establishing a cluster-retrieval *upperbound* - the best effectiveness that could (theoretically) be achieved with a certain clustering method.

Inverted-file based - If the data structure stores, for each vocabulary term, all the clusters for which the term is highly topical (the term is in the cluster representative), then a query-based search can rank the clusters based on their estimated relevance to the query. Such an inverted file can be built when the cluster representatives in the cluster structure are computed.

2.3.10 Cluster representative

The need for a *centroid* or *cluster representative* that summarizes and conveys the content of the cluster comes from the need to compare two clusters (typically in hierarchic clustering), a document with a cluster (typically in non-hierarchic clustering) or a query with a cluster (when searching). For visualization tools based on clustering, it is based on the cluster representative, or *label*, that the user may decide whether a cluster looks promising and is worth exploring.

One approach to representing a cluster is to select one or several actual documents of the collection, most representative for the cluster. This is usually the option of choice when the media of the documents does not afford summarization. For text documents it is more common to summarize the document contents into a *cluster representative*, which contains the terms most specific to the cluster.

Several versions of an *un-weighted representation* were proposed by Jardine and van Rijsbergen [JR71]. Of these, the label that was most successful in search effectiveness experiments was the one that included terms present in at least $\log_2 |C|$ documents, $|C|$ being the number of documents in the cluster.

A *weighted representative* can also be used, as described by Voorhees [Voo85b]:

- The sum of the within document frequency of each term in the cluster is computed,

and the terms are sorted by decreasing frequency.

- The top terms are selected to be in the centroid. The weight of each term in the centroid is the rank (from the bottom) of the term in the sorted list; equal frequencies are assigned the same rank.

- The rank weights are multiplied by an inverse document frequency factor (over the collection) and normalized so that the sum of the squares of the obtained weights equals one (cosine normalization).

More recently, models and formulae that proved successful in indexing documents in view of best-match searching have been applied to generating cluster representatives. For example, Neto et al. use the *bag of words* as common representation for documents and clusters, and the classic tf-idf formula for weighting and ranking the terms in order to produce labels for visualization [NSKF00].

As in the case of document representatives, built through indexing, the cluster representative must give a balance between *representativeness* (or accuracy) and *discrimination power*. More index terms give a more accurate representation of the cluster, but it is more difficult to discriminate between clusters. With fewer index terms, cluster representatives are more easily distinguishable, but they represent the cluster less accurately. Of course, the representativeness and the discrimination power of the cluster representative are determined not only by the actual terms, but also by their weights.

If cluster representatives are used for browsing in an interactive environment, the particular visualization tool may impose constraints with regards to the size of the label. If the user interface allows for space, then the content of a cluster can be represented not only by *topical terms*, but also by *typical document titles*, as Scatter-Gather does [CKPW92].

It is worth mentioning that, as the cluster representative is, conceptually, similar to a document representative, the similarity measures available for documents can also be used for the query-cluster match. The most widely used measure is the *cosine coefficient*, but *probabilistic matching functions* have also been described by Yu [YL77] and used by Salton and Wong [SW78].

2.3.11 Evaluation of clustering

Clustering was introduced and adapted to IR as a means to structure a document collection in order to increase the efficiency and effectiveness of retrieval. In trying to estimate the success of the clustering approach, various evaluation studies have looked in three main directions trying to measure:

1. the classifiability of a document collection i.e. its intrinsic capacity to be semantically structured based on the topics and subtopics that it covers.

2. the capacity of clustering methods to produce a meaningful structure.

3. the capacity of a structure obtained through clustering to support a retrieval process.

Most researchers did not distinguish between the first two issues above, so they are reviewed together in the next subsection. The following subsection looks at the third issue, the evaluation of cluster-based retrieval, which has been the main type of evaluation done on clustering. More recently, clustering has been used to guide exploration, so a subsection is dedicated to that subject. However, to date no evaluation of clustering as a tool to support exploration has been proposed. That issue is addressed later in the thesis.

Evaluation and validation

The two concepts are strongly related and therefore often used interchangeably: evaluation measures the quality of clustering, while validation compares it to a prescribed or expected quality level.

Studies on evaluation and validation of clustering attempt to answer questions such as the ones proposed by Dubes and Jain [DJ79, JD88]:

- Which clustering method is appropriate for a particular data set ?

- How does one determine whether the results of a clustering method truly characterize the data ?

One questions that has not been given much attention is:

- Which clustering method is appropriate for a particular task ?

This is probably due to the fact that traditionally document clustering has been seen as a tool with rather limited applicability in IR. Most studies have evaluated it in the context of cluster-based retrieval in batch mode.

In his thorough, even if somewhat dated, review of clustering [Wil88], Willett identified three main classes of evaluation:

1. **Theoretical studies** attempt to provide a mathematical analysis of the characteristics of various methods, assessing if they satisfy certain *criteria of adequacy* [Rij79] such as:

   - The hierarchy is stable under growth.
   - The method is stable under small errors.
   - The method is independent of the initial order of objects.

   The *single-link* clustering method is by far the best from this point of view. Actually it is the only one among the HACMs described that satisfies all these criteria.

2. **Simulation studies** involve the generation of an artificial data set so that the true structure of the data is known. What is tested is the *recovery characteristics* of the methods and their *stability* under various error conditions are studied, as well as the difference between the structure obtained through clustering and a structure generated randomly. The *Ward* and *group average* methods tend to give the best performance while the *single-link* performs consistently poorly [MSS83].

3. **Practical studies** use real data and attempt to evaluate clustering on the basis of the usefulness of the classifications produced (such as effectiveness of retrieval, grouping together of documents relevant to each of a set queries for ...) or by comparison with manual classifications of data. The *single-link* method performs consistently poorly in terms of retrieval effectiveness, compared to the other HACMs. The *complete-link* method, combined with top-down search, usually gives good effectiveness [Voo86].

The experimental evaluation of clustering algorithms has shown that the results are highly dependent on the actual collections on which clustering is applied, so the effects

of the clustering and of the collection may be confounded. Several tests have therefore been proposed to measure the *clustering tendency* or *classifiability* of a collection, which indicates if a document collection is likely to respond well to the application of a clustering method:

The cluster hypothesis test, or *separation test*, due to Van Rijsbergen and Sparck Jones [RSJ73, Rij79], is based on the hypothesis that documents similar to each other are expected to be relevant to the same queries and that dissimilar documents are unlikely to be relevant to the same requests. The test calculates all the relevant-relevant (RR) and relevant-nonrelevant (RNR) inter-document similarities and checks that the average RR coefficient is larger than the average RNR coefficient. Furthermore, the coefficients can be summed over a set of queries and plotted as relative frequency histograms in order to check the overlap of the two distributions. The less overlap, the better separation between relevant and nonrelevant documents.

The nearest neighbour (NN) test, proposed by Voorhees [Voo85a], attempts to address the distortion caused, in the separation test, by non-relevant documents having a much higher relative frequency than relevant documents. The NN test takes in turn each of the documents relevant to a query and identifies how many of its nearest neighbours are also relevant to the query and sums this over a set of queries. A high percentage of relevant documents that have other relevant documents as nearest neighbours is seen as a confirmation of the cluster hypothesis.

The term density test, due to El-Hamdouchi and Willett [EHW87] looks at the average number of terms that index each document of a collection compared to the size of the vocabulary. A high depth of indexing is expected to better differentiate between documents which are highly similar and those which are more distantly related, this resulting in a highly structured classification, with clusters of very similar documents nested within larger clusters of a more heterogeneous nature. The authors claim that their experiments showed this test to be more accurate in measuring the clustering tendency of a document collection than the previous two. Moreover, it is faster and does not require relevance judgements, like the previous methods.

However, as the experiments described by the test's authors only considered binary representation of documents and their model assumed an even distribution of index

terms over the documents, we question whether their results can be generalised.

Evaluation of cluster-based retrieval (CBR)

Traditionally, cluster-based retrieval was seen as the retrieval process based on a clustered collection which retrieved one cluster estimated to best match the user's query. In Web-Cluster, a highly interactive system based on the exploration of the structured collection, we have relaxed this view and instead rank the clusters based on their estimated relevance to a query. The best cluster is still retrieved, at the top of the ranked list, and additionally the user has the opportunity to explore other good clusters.

In discussing the evaluation of CBR we first review the traditional approach and then comment on ways of extending it in response to the more modern, interactive approach. In order for this section to make perfect sense, the reader is expected to have an understanding of evaluation of IR systems. If that is not the case, then section 2.4 should be read first.

Traditional CBR

Early attempts to evaluate cluster-based retrieval took place in the evaluation framework of the time: a system approach based on test collection, with a set of constructed queries, and associated relevance judgements, produced by human experts. The human factor was not taken into account.

The two aspects that need considering in such an approach are *efficiency*, which looks at the resources needed in order to obtain a result, and *effectiveness*, which looks at the quality of the result.

A machine independent approach towards evaluating *efficiency* of CBR is the analysis of the algorithms used for the initial indexing and clustering of the collection and for searching the cluster structure. Although this gives an idea of the algorithms complexity and of the expected performance, the absolute performance is influenced by different optimisations like upperbounds, similarity matrix reduction, efficient use of an inverted file, and by other factors specific to a collection, like richness of indexing (or *indexing*

exhaustivity). Consequently, the efficiency of retrieval for a certain collection on a certain machine, is most often given by the computer resources used such as storage and processing time. In any case it should be measured in conjunction with the retrieval effectiveness in order to give an idea of the cost per quality.

The *effectiveness* of retrieval is commonly measured in terms of *recall (R)* (the ratio of relevant documents that are retrieved) and *precision (P)* (the ratio of retrieved documents that are relevant). When relevance information is available, typically for a test collection, R and P associated to each cluster can be used as measures of cluster quality. In order to allow flexibility and to consider both recall-oriented and precision-oriented tasks, and to provide a unique measure of quality based on which clusters can be ranked, the *effectiveness measure (E)* has been introduced [JR71]:

$$E = 1 - \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R} \tag{2.12}$$

The parameter β defines the relative importance the user attaches to recall compared to precision: $\beta = 1$ attaches equal importance to recall and precision, while a higher β biases the E formula towards recall and a lower β towards precision. E takes values between 0 and 1, with low values indicating good effectiveness ($E = 0$ for *complete success* and $E = 1$ for *complete failure*), and is normally calculated for $\beta = \frac{1}{2}$, $\beta = 1$ and $\beta = 2$.

This function is the *de facto* measure of effectiveness in cluster based retrieval[7], although some researchers prefer its complement $F = 1 - E$ [SJBH97]. It is apparent that F can be viewed as a biased harmonic mean:

$$F = \frac{\beta^2 + 1}{\beta^2 \cdot \frac{1}{R} + \frac{1}{P}},$$

and for $\beta = 1$, F becomes the harmonic mean of R and P:

$$F = \frac{2}{\frac{1}{R} + \frac{1}{P}}.$$

Various experiments have tried to 'rank' the clustering methods in terms of retrieval

[7]It can also be used for comparing cluster-based retrieval with best-match retrieval in terms of effectiveness.

effectiveness that they produce. The only consistent result is that the *single-link* method performs poorly compared to the other HACMs. The results of comparing the other HACMs vary widely, the conclusion [Wil88, Bur95] being that the effectiveness of cluster-based retrieval depends on the test collection, on the data representation and especially on the indexing exhaustivity, on the inter-document similarity measure, on the search strategy, and on various other parameters.

One approach is to ignore the search method and to concentrate on the quality of the clustering, quantified by different measures of optimal retrieval performance, assuming that future search strategies may approach this performance. Another, opposite approach is that clustering is not an end in itself but a tool for supporting retrieval and that it is performance of existing search strategies that should be compared for a certain cluster structure. The choice depends on the use of clustering.

Interactive CBR

In recent years, the feeling has grown that recall-precision evaluation is inadequate for modern, interactive retrieval systems, and alternative methods have been investigated. While no method has gained widespread acceptance, the tendency seems to be towards a user-centred, task-oriented approach, that investigates retrieval as part of a complex task or problem solving system, in an information seeking environment.

These investigations are in their infancy, but the results are encouraging. Experiments with Scatter-Gather, a system that partitions search results, has shown that clustering helps the users better grasp the topical structure and the vocabulary of a domain [PSHD96], and also that using clustering can significantly improve effectiveness of a system based on ranked retrieval [HP96].

2.3.12 Current trends in clustering

Cluster-based retrieval was initially proposed as a more efficient and potentially more effective alternative to ranked retrieval. However, despite the amount of research that went into developing methods, algorithms and similarity measures in order to increase its effectiveness, CBR did not live up to the expectations that it had raised:

- No clustering method has proved consistently (i.e. on the majority of test collections) superior to best-match searching in terms of effectiveness [Wil96].
- Modern query expansion techniques have offset CBR's advantage of identifying relevant documents which do not match the initial query.
- Modern search algorithms based on inverted files have become quite efficient, plus they do not impose the same overhead as clustering, expensive in terms of disk space and processing time [Voo85b].
- The ever-growing size of test and operational collections has made clustering inadequate as a tool for structuring full document collections.

At the time when the confidence in clustering as a tool for information retrieval was decreasing, the 'interactive revolution' was taking place. IR researchers were coming to accept that IR systems should be highly interactive and should support the searcher in various stages of the information seeking process: problem definition, source selection, problem articulation or examination of results. This gave clustering a second chance, as it has come to be perceived as a tool appropriate for structuring sets of documents in view of exploration.

Scatter-Gather [CKPW92] proposed the use of *browsing* as the main information access paradigm and was one of the first interactive IR systems to promote the use of clustering as an adequate tool for organising search results. In Scatter-Gather clustering was part of the iterative search interaction: applied either to relatively small collections, or to the output of an initial, vague search, a partitioning algorithm *scatters* the documents into clusters; the user browses the clusters and *gathers* those that, according to their label, seem promising, thus narrowing the information space to be explored. A combination of browsing of the information space based on a table-of-contents paradigm, selection and gathering of the promising information sub-spaces, and re-partitioning of the selected set of documents is expected to help the user explore the problem domain and identify relevant documents.

Browsing has been shown to be particularly useful when the user is not looking for anything specific, but rather wishes to get a feel for the topics of a collection or hopes to

serendipitously find something interesting [Mar95]. Obviously, browsing can be combined with searching: browsing can help refine a query, which is then used for a search; on the other hand, searching can provide starting points for browsing.

Browsing information spaces has proven to be popular, to the extent that some researchers have envisaged the creation of query-less information access systems: the potentially user-unfriendly query formulation mechanism is eliminated and instead visualization tools that convey the structure and content of the domain are offered for the user's exploration of the document space [Dub95, Dub96, CR96], sometimes accompanied by tools for indicating relationships between documents or for filtering documents based on various attributes [Kor91, Spo94].

With the advent of the Web and the building of *digital libraries* the size of the collection on which retrieval is done has increased dramatically, so clustering full collections in view of supporting retrieval may be difficult or even impossible. The most common approach at the time of writing seems to be [ZEMK97, HP96] the *query-specific clustering* approach proposed by Willett [Wil85]: a 'traditional' (Boolean or best-match) search, followed by the clustering of the subcollection returned, containing only the documents considered relevant to the query. A consequence of this is that the clustering is done on-the-fly, so fast clustering algorithms need to be applied, often with a trade-off in retrieval effectiveness.

2.4 Evaluation and experimentation in Information Retrieval

2.4.1 Introduction

This thesis proposes a new interactive model for Information Retrieval, system-based information access based on structured specialised collections, and conjectures that this model is expected to improve retrieval effectiveness. In order to support this claim, the appropriate evaluation needs to be conducted.

The purpose of this section is to review evaluation methodologies used in IR so that the appropriate evaluation framework can be constructed. The two questions that guide this review are:

1. What to evaluate ?

2. How to evaluate ?

No fundamentally new algorithms are proposed in this thesis, so we can ignore issues of efficiency, as long as the user interaction is not affected. We concentrate mainly on effectiveness of retrieval and on user satisfaction.

2.4.2 IR evaluation

There are two fundamentally different and complementary approaches to evaluating IRS. The *systemic* approach takes a narrow view of the definition of an IRS, limiting it to the function of indexing and retrieving documents. The user's contribution is practically ignored, the experimental setting considering a fixed user profile, represented by a fixed set of information needs. This is the classic approach and is still used when the purpose is to improve representation models, data structures, or algorithms. In this case one can distinguish between black-box or diagnostic experiments. In the former case the system is treated as a whole, and the output is observed for a certain input. In the latter case it is the internals of the system that are observed, and their influence on the output [RHB92].

A more modern, *user-centred* approach was brought about by a combination of three 'revolutions': the relevance, the cognitive, and the interactive revolutions [RHB92]. An IR system is most often viewed as an interactive system that should allow the user to explore a problem domain and to gather relevant information identified through a combination of browsing and search strategies. The system is put in a broad perspective, as an *information seeking environment* [Mar95, HH96, Hen96, HH97] that supports the user in planning search tasks, developing a search strategy, retrieving and organising information, and possibly monitoring the state of search tasks [FHH96]. Actually even the term "information retrieval" is often replaced with "information access", indicating that the process is of exploring the information space rather than retrieving bits of information. The user, whose information need is seen as an *anomalous state of knowledge* (ASK) that needs to be resolved [BOB82] is at the centre of the evaluation setting. The user's knowledge and, consequently, the information need may change during the interaction with the system, and so may the relevance of the retrieved documents, in the view of the

information/knowledge acquired by the user.

A systemic view can be taken to evaluate some algorithms or components of an IR system, and a 'strictly interactive' view can be used when under investigation is the usability of a system and the user's interaction with the system. However, neither view, taken in isolation, can give an overall view with regards to the usefulness of a system for a certain task. The best retrieval algorithm is ineffective if the interface to it is unusable; conversely, a clear and intuitive interface is not worth much without a good retrieval engine behind it. Therefore, the design of effective interactive retrieval environments will require careful attention to the larger *human - interface - retrieval engine* system and a complete evaluation should look at its capacity so solve the types of tasks it was designed for [BBDHB94].

2.4.3 Traditional performance measures in IR evaluation

Efficiency and *effectiveness* are, usually, the measures considered when evaluating an IRS. The first measure usually looks at the time and space requirements of the algorithms used by the system, checking whether operations such as indexing, storing the index, searching, or clustering are possible or acceptable in terms of functionality. This is especially important for very large collections that need indexing and searching, such as the Web, and for interactive systems, for which the search time needs to be in the order of seconds.

While efficiency acts like a filter, confirming or not the viability of a system to a task, it is usually the effectiveness that is considered the measure of retrieval quality, measuring the system by the quality of its results. The preferred experimental methodology is the *laboratory* setting, based on *test collections*, which comprise a document collection, a set of information requests and, for each request, the set of documents considered relevant by domain experts. The classic measures calculated are *recall*, i.e. the fraction of the relevant documents which has been retrieved, and *precision*, i.e. the fraction of the retrieved documents which is relevant. Details on these measures, as well as others (often derived from recall and precision), are presented in [BYRN99]. Also see [Hul93] for considerations regarding statistical methods in IR experiments.

While a variety of test collections were created and used in tests by different research groups, the need for a common test collection, of realistic size and content coverage, a common task environment and common methodology, became apparent in order to support comparisons among systems and techniques, and ultimately, to support progress in IR research. The solution was TREC, an international benchmarking exercise where systems are tested in a common environment consisting of a number of document collections, a set of *topics* (descriptions of information needs) and relevance judgements that mark the relevant documents for each topic [Har93]. An important advantage is that the environment and the design methodology for a variety of typical information seeking scenarios have been developed in common by the research community, so it is accepted as a sound experimental setting, despite critics regarding its limitations and the validity of its assumptions [Sar95]. Of course, some flexibility is allowed, so that aspects deemed important for particular systems can be evaluated, apart from the common evaluation.

The single-valued evaluation measure preferred in TREC is the *average uninterpolated precision (AUP)* [VH00]. The average precision for a single topic is the mean precision obtained after each relevant document is retrieved (using zero as precision for relevant documents that are not retrieved). The average precision for a run is the mean of the average precision scores of each of the individual topics in the run. It has a recall component in that it reflects the performance of a retrieval run across all relevant documents, and a precision component in that it weights documents retrieved earlier more heavily than documents retrieved later.

2.4.4 Interactive IR evaluation

The traditional, batch-mode experiments (in which a batch of queries are submitted to the system, with no user intervention or parameter adjustment in-between) are good for testing and improving models, formulae, algorithms, but do not reflect the real use of a retrieval system in an interactive environment. Some aspects that indicate their inappropriateness for interactive systems are:

- Document relevance is a complex human cognitive and social phenomenon, dependent on circumstances and context [Sar95], so even the experts introduce bias when making relevance judgements.

- The relevance of documents, rather than being static, may be dependent on the user's task, and on the documents already seen by the user.
- Real queries formulated by real users (as found in search engine logs) are much shorter than the ones derived from the artificially built topics in test collections.
- Users tend to examine only the top-ranked documents retrieved by the system, so measures such as recall and precision at point 200 or 1000, typically used in batch tests, are irrelevant.
- The response time is important for the user, while it is generally ignored in effectiveness evaluation.

A new type of evaluation is needed that takes into account the user and her interaction with the system. A difficult issue is the choice between *laboratory* versus *operational system* tests. The conflict is essentially between, on the one hand, control over experimental variables, observability and repeatability, and on the other hand, realism. Real users' behaviour and the relevance and utility of retrieved documents for real users are extremely difficult, if not impossible to simulate. Therefore, tests with real users and real tasks in an operational environment are needed in order to validate a system. New measures of quality, such as user satisfaction (based on content, accuracy, format, ease of use, and timeliness [DT88]) or task-oriented measurements such as success, completeness, time, cost, utility and so on can be employed.

This kind of operational tests are less appropriate if the purpose of the experiment is to compare the effect of various parameters to the quality of retrieval, to compare alternative formulae or alternative solutions for a certain system component, or to make objective estimations regarding the quality of the system. In this case a better approach is a controlled experimental design that reveals the effect of the chosen parameter on the retrieval effectiveness and compensates for variation in other parameters (including the user).

Probably what is needed in order to establish the parameters that give best performance, and also to make sure that the system is usable and useful, is a combination of laboratory and operational tests, so that the effect of various components of a system is

investigated under controlled conditions, and a generic, or alternatively an optimal version of the operational system is tested with real users [RHB92].

Borlund and Ingwersen have taken the middle ground by proposing the use of *simulated work task situations*: real users are immersed in realistic scenarios and are assigned tasks [BI97]. Apart from this context restriction, which constitutes the commonality between users, the experiment setting is realistic in the sense that the users derive their own information need, formulate queries, examine results, reformulate queries, and attempt to solve the task assigned. Even the context restriction is quite realistic, as it is common for information needs to appear from work assignments or in other social environments, rather than from the user's own interest. Such an experiment can be used in order to study the user's behaviour during the information seeking approach, mental models and search strategies, the system's usefulness and usability with regards to achieving the assigned task, the user's general satisfaction with the system. The two authors have shown that there is no significant difference in searching behaviour for real, respectively simulated information needs[8] [BI99]. This is a very useful result which opens the way for more extended laboratory tests based on simulated work task situations: if realistic scenarios can be built on realistic information needs, then more objective measures, such as the effectiveness of the system, can be used (either by asking the users to mark the relevant documents, or by using expert assessors).

A step in that direction is taken by Reid who proposes the use of a task-oriented test collection and what she claims is new evaluation methodology, centred around the task [Rei00]. The relevance of a document is not intrinsic, in this case, but determined by the contribution of the document in completing a certain task. It is debatable, however, whether Reid's task-oriented approach is fundamentally different from TREC's topic-oriented evaluation as long as a typical task is to gather information on a certain topic.

Probably the most serious challenge in evaluating interactive IR systems and a realistic situation is posed by its most complex and least understood 'component': the *user*. A quick look at the variables in an interactive system, as identified by Hersh [HR97], indi-

[8]The only exception is that users took longer to read documents in the case of real needs, probably due to a higher personal interest.

cate that user-oriented measures are more numerous, more difficult to measure or estimate (both input and output) and more difficult to control (input).

This complex variability has been addressed in two complementary ways. On the one hand is a painstaking set of experiments that introduce incremental, small variations, trying to identify various aspects of the interaction [Bel98]. On the other hand is a combinatorial design that attempts to statistically eliminate the effect of some variables and to identify the effect of others [LO98]. User-centred interactive retrieval evaluation is still in its infancy, but it is hoped that it will give a better understanding of the behaviour of the user during the information seeking process and will contribute to designing better systems.

2.4.5 The Interactive track of TREC

Although TREC was originally only designed to support laboratory testing, in order to compare and improve system components, it was soon criticised for failing to take into account the human contribution to the retrieval process. Therefore, participants were allowed to submit results obtained by using interactive queries (queries developed by human searchers while interacting with the system) in addition to results obtained based on automatic (queries resulting from fully automatic processing of the query) or manual (queries whose generation involves some human intervention, but without interacting with the data) ones.

An analysis of the poor results obtained by the interactive queries compared to the automatic queries, revealed some inadequacies of the experimental setting and design [BRR96]. Firstly, the professionals who constructed the topics had access to the data, so the topic descriptions could be quite precise and specific, while in reality the user's information need is usually not clearly defined from the outset. Secondly, the influence of the search intermediary was evident in the description of the topics, easy to translate into search queries. Moreover, it was recognised that, compared to an operational setting, the experimental setting for these interactive experiments was not realistic, as

- the user's actions were too restricted,
- the queries were not realistic,

- the high number of queries not only made it difficult to find users, but it added the effect of attention and tiredness,
- the experimental design was not appropriate, as repetitions are difficult, expensive and unlikely to produce identical result in an interactive setting.

The need for specific experimental setting (including document set, queries, experimental design) for evaluating interactive systems was recognised in the TREC community and an Interactive track was created[9]. The initial attempt at compatibility with the main track was soon dropped and an independent experimental design for the Interactive track was proposed. Of course, a laboratory experiment is unlikely to perfectly match the real situation (so complementary, HCI-specific experiments are encouraged), but the TREC-like experiments are accepted by the IR community as satisfactory for evaluating an interactive system and the search process. Its recognised inadequacies have been addressed by the continual review and the alterations to the design done by the participating researchers.

It is not an objective of this thesis to revolutionize evaluation of interactive retrieval, so the design of the experiments will owe a great deal to the Interactive track of TREC, with the details determined according to the specificities of WebCluster, its target users and the tasks expected to be solved by it. We will have, therefore, a brief look at the design of TREC interactive experiments and its evolution.

The interactive track adds a new variable, the user, to the variables considered by the other experiments, the system and the topic. A balanced, Latin-square design is needed to compare the effects of each variable. Ideally all combinations of factors, with repetitions, should be considered, but this is not feasible within a participating site (research group) and certainly not across sites. For example, a user cannot perform a search for a topic more than once, because the learning effect would bias a second search. Moreover, reliably detecting significant system effects requires relatively many searches. A solution would be to use a high number of users and randomly assign them to searches on different systems, at different sites. For logistic reasons, this is, however, impossible.

[9]The reader interested in the evolution and details of the TREC Interactive track are directed to [Ove01].

The approach adopted by NIST in TREC-6 (1997) was to compare the experimental system with a baseline system: better retrieval results would show better support for the user in retrieving relevant information [LO98]. NIST offers its own system, Zprise[10] as a general baseline system, so that indirect comparisons can be made between various systems. The document collection used was Financial Times of London, 1991-4 (with 210,158 articles totaling 564 MB), a sub-collection of the one used for the ad-hoc track, and the topics were modified topics from the main track. The searcher's task was to find and save documents that taken together contained as many answers as possible to the questions stated in each of a set of 6 test topics. Recall and precision was measured in terms of all possible answers as determined by NIST assessors.

The novelty in TREC-7 and TREC-8 was the task of finding not documents, but as many as possible relevant aspects (or instances) of the answer to each topic. The effectiveness of the search was evaluated by the fraction of total instances for that topic that were covered by the search (*instance recall*) and the fraction of the documents retrieved that contained an instance (*instance precision*). This was motivated by the need to investigate a different aspect of the search process: finding information rather than documents. Relatively complex topics were selected in order to make the search for aspects feasible. Higher importance was given to recall than to precision: searchers were encouraged to avoid saving documents which contributed no instances to the documents already saved, but there was no scoring for saving such documents and the searchers were told that.

For TREC-9 the intent of the Interactive track was to explore tasks similar to those common on the Web: finding answers to relatively short queries. The intent was to also use a Web sub-collection, i.e. a sample from the World Wide Web in order to have a realistic collection in terms of size and type of documents, and to offer the participants the possibility to make use of hyperlinks. In the end, the collection was not available, the track becoming an interactive version of the Question/Answer track. The measures of effectiveness were precision and relative recall. Measures such as elapsed clock time and user satisfaction were also taken, and supplementary statistical analysis of the data was

[10]http://www-nlpir.nist.gov/works/papers/zp2/zp2.html

encouraged.

For TREC-10 (2001) the general consensus was to investigate interaction in Web searches. The experiment is conducted in two stages, over 2 years. In the first year an *operational* setting was used, allowing the study of user behaviour and, as a by-product, building a collection of Web documents, with user relevance judgements. These will be used in the second stage, when quantitative measures of effectiveness will be measured in a *laboratory* setting.

2.5 Conclusions

Our review has identified some gaps in Information Retrieval research which are going to be addressed in this thesis. Let us go over these areas.

- **The interaction model**

  Although the use of relevance feedback for query reformulation has been thoroughly investigated, there has been little work on interaction models that support mediation, i.e. system support for the user in conveying an information need and exploring a problem domain. We intend to simulate the interaction that takes place in the library and to have the system emulate the librarian in eliciting information from the user and in guiding the user's search.

- **The use of document clustering**

  While initially clustering was used mainly for cluster-based retrieval (CBR), most current uses of clustering are for organising search output. We accept that CBR on its own is unlikely to significantly improve effectiveness of retrieval, so we propose clustering as a means for structuring specialised document collections in view of supporting exploration of various problem domains. Combined with appropriate information visualization tools, the collection structure can support a combination of browsing, best-match and cluster-based searching, and can potentially yield improved effectiveness and user satisfaction.

- **The cluster hypothesis**

  Although the cluster hypothesis has been shown to hold on search results and to be useful for organising search output, results on clustering full collections have been inconclusive. Following quite recent experiments, some authors have dismissed clustering as an effective tool for supporting retrieval, concluding that the cluster hypothesis is unlikely to hold, as it is based on the wrong assumption that topical relatedness is equivalent to a relevance relationship [SJBH97]. We consider this conclusion to be inappropriate: the statistical approach to Information Retrieval is based in its entirety on the fundamental assumption that the words in a document represent its content. The failure of the cluster hypothesis, when it fails, is not intrinsic to clustering, but is due to the failure of automatic indexing to identify the *aboutness* of a document, i.e. the topic and concepts that the document is about.

  We intend to conduct an investigation on the cluster hypothesis and to refine its original formulation.

- **Cluster representatives and topic models**

  Traditionally, heuristics have been used to generate cluster representatives, in a process independent of the indexing of documents. We intend to integrate the two processes into a topic model framework based on probabilistic language models. The user's topic of interest may be represented by individual documents, or by a set of documents grouped in a cluster. It is desirable to propose a method to generate a topical model (i.e. a representation of the user's topic of interest) in either case.

This review has also provided a good understanding of the tools, techniques and methods that can be employed for investigating and filling in the gaps identified. By concentrating in the specific areas that support our approach to mediating retrieval we attempted to convey our choices of models and tools:

- A weighted vector representation of documents.

  While only term frequency information is stored permanently, the actual weights are computed based on the chosen weighted scheme when clustering or searching is performed on the document collection.

- Hierarchic agglomerative clustering methods.

  A clustering framework has been implemented, so an experimental system can use and compare a variety of methods. The one that is going to perform best in a certain context will be used in the operational system.

- Exploration tools based on visualizing hierarchic structures.

  We have implemented a framework for building various user interfaces, which support different mediation scenarios. These interfaces are all based on visualization tools for exploring the source collection and have been designed to work equally well with clustered or manually classified collections.

- TREC-like evaluation.

  For testing our ideas we have built an evaluation framework inspired by and using test collections from TREC. However, the code for testing is flexible and rather generic, so the experiments can be easily repeated on different document collections (as long as the same kind of relevance judgements are available) or adapted.

The rest of the thesis discusses in detail the mediated retrieval concept and tests its assumptions and claims.

Chapter 3

System-Based Mediated Access and Contributions to Research

3.1 Introduction

The first two chapters introduced the concept of system-based mediated retrieval and reviewed models, technique and evaluation frameworks that we can adapt and use to implement and test it. Here we describe the mediated access concept in more detail and present the objective of the thesis as well as the intended contributions to IR research.

Firstly, we describe the interaction model of mediation and the metaphor it is based on, discuss theoretical aspects as well as implementation decisions, and propose some practical applications. Secondly, we propose a set of research hypothesis and an evaluation framework for testing them.

3.2 Discussion of the mediation concept

3.2.1 The mediation process in more detail

The concept of *system-based mediated information access*, proposed in this thesis, refers to the system assisting the user in investigating a domain of interest, in exploring and refining an information need, and in generating a query that conveys the information need accurately.

Our approach to mediated information access is based on the existence of *specialised collections* of documents or abstracts maintained by various companies and organisations, such as MEDLINE for medicine, CABI for agriculture and biosciences, or RAPRA for the polymer industry. These collections are kept up to date in terms of validity of the information and comprehensive coverage of various specialised domains. The task they typically support is searching for information in that particular domain. We propose to use these collections, representative for the domains they cover, as 'source collections' for mediation.

In order for the user's exploration of the problem domain to be possible, the source collection needs to be structured. Any classification method, either manual or automatic, will do, as long as the topical structure is revealed and can be exploited by information visualization tools and by a combination of searching and browsing strategies. Some of these specialised collections are already categorized by their creators, typically based on a manually created taxonomy of the domain.

In this thesis we intend to investigate the capacity of clustering to group similar documents, to reveal the topical structure of a specialised collection and thus to support exploration. There are several reasons for our choice. Clustering is a fully automatic procedure, and therefore fast, cheap, and domain independent. Its parameters (such as the choice of clustering method, similarity measure and various thresholds) can be changed easily, so their effect can be compared. Clustering is applied after a collection has been indexed; therefore, it allows for flexibility in using various indexing parameters and weighting schemes, providing more ground for research and experimental results.

The mediated retrieval process is depicted in **Figure 3.1.** Tools are offered to the user for exploring the structure, the topics and the terminology of the domain, represented by the specialised source collection, thus supporting a learning process for the user unfamiliar with the domain. Moreover, the user is invited to explore the use of tools that implement various retrieval strategies, and can therefore learn what strategies are available, which ones are more appropriate in a given situation, and how they can be combined. Based on the user's exploration of the collection, and on her selection of relevant documents, the system builds a *(statistical language) model* of the topic investigated. It can then act as

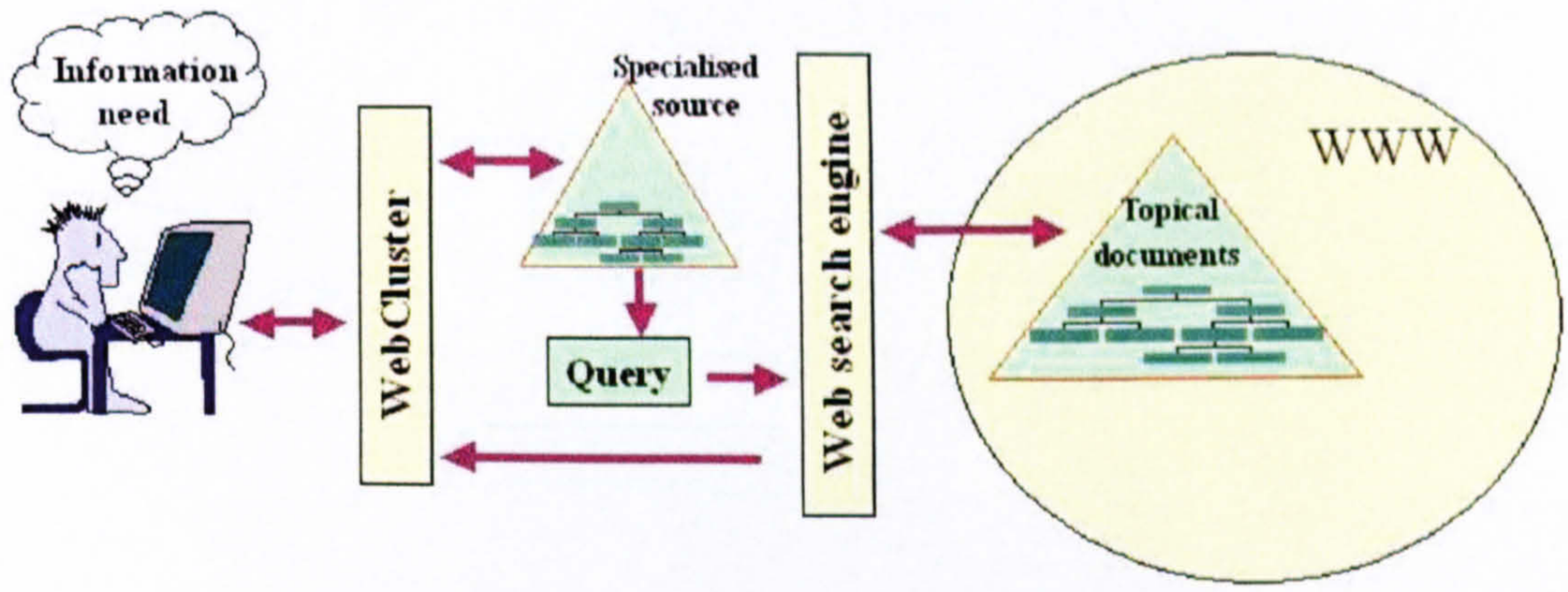


Figure 3.1: Mediating access to the World Wide Web.

a *mediator* by generating a query that comprehensively, clearly and precisely reflects the contents of the documents selected by the user. This *mediated query* can then be used to extend the search to any 'target collections' that are heterogeneous, unstructured and too large to readily afford exploration strategies other than query-based searching, such as the World Wide Web. We conjecture that mediation through the right 'source collection' has the potential to generate a very precise query and to significantly increase the quality of the retrieval effectiveness and the perceived completeness of the user's task.

The source collection acts as a filter for the target collection: the mediated query, built based on source documents, will retrieve similar documents from the Web. Moreover, the user can explore various topics of the specialised domain and generate a series of mediated queries. Therefore, the structure of the specialised domain is conceptually projected to, or imposed on the target collection.

Figure 3.2 shows the UML diagram of the simplified mediation retrieval process. It is apparent that the system has two distinct operation modes, for the two stages of document retrieval. In the first, *exploratory* stage, the system supports the user's exploration of the domain of interest and the formulation and refinement of her information need. In the second stage of mediated retrieval, the system is in *search mode*: the mediated query is submitted to the so-called 'target collection' and high retrieval effectiveness is sought.

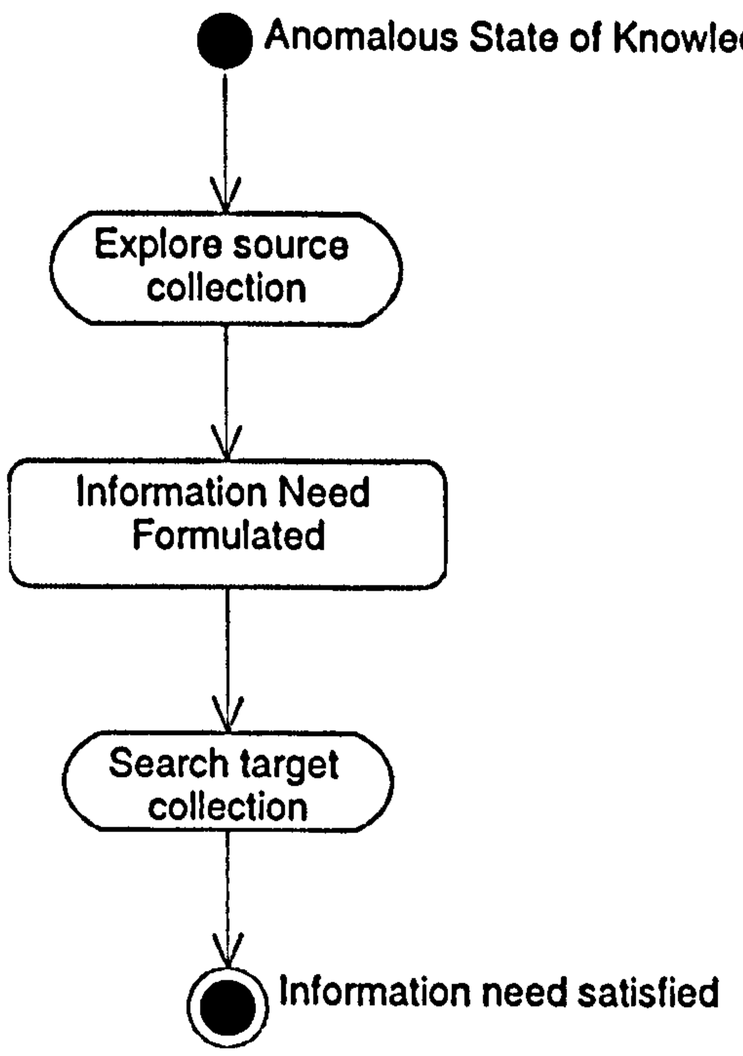


Figure 3.2: Simplified mediated process.

A closer look at mediation, as shown in **Figure 3.3**, reveals that a rather complex process takes place. Starting with an information need represented by an *anomalous state of knowledge*, the user can employ a variety of search strategies in order to explore the source collection representative for the domain of interest:

- ranked searching
- cluster-based searching
- browsing of the hierarchical structure
- a combination of the above

The user can choose to expand clusters or classes of documents for further exploration and can choose to display and read documents, thus ostensively indicating her information need.

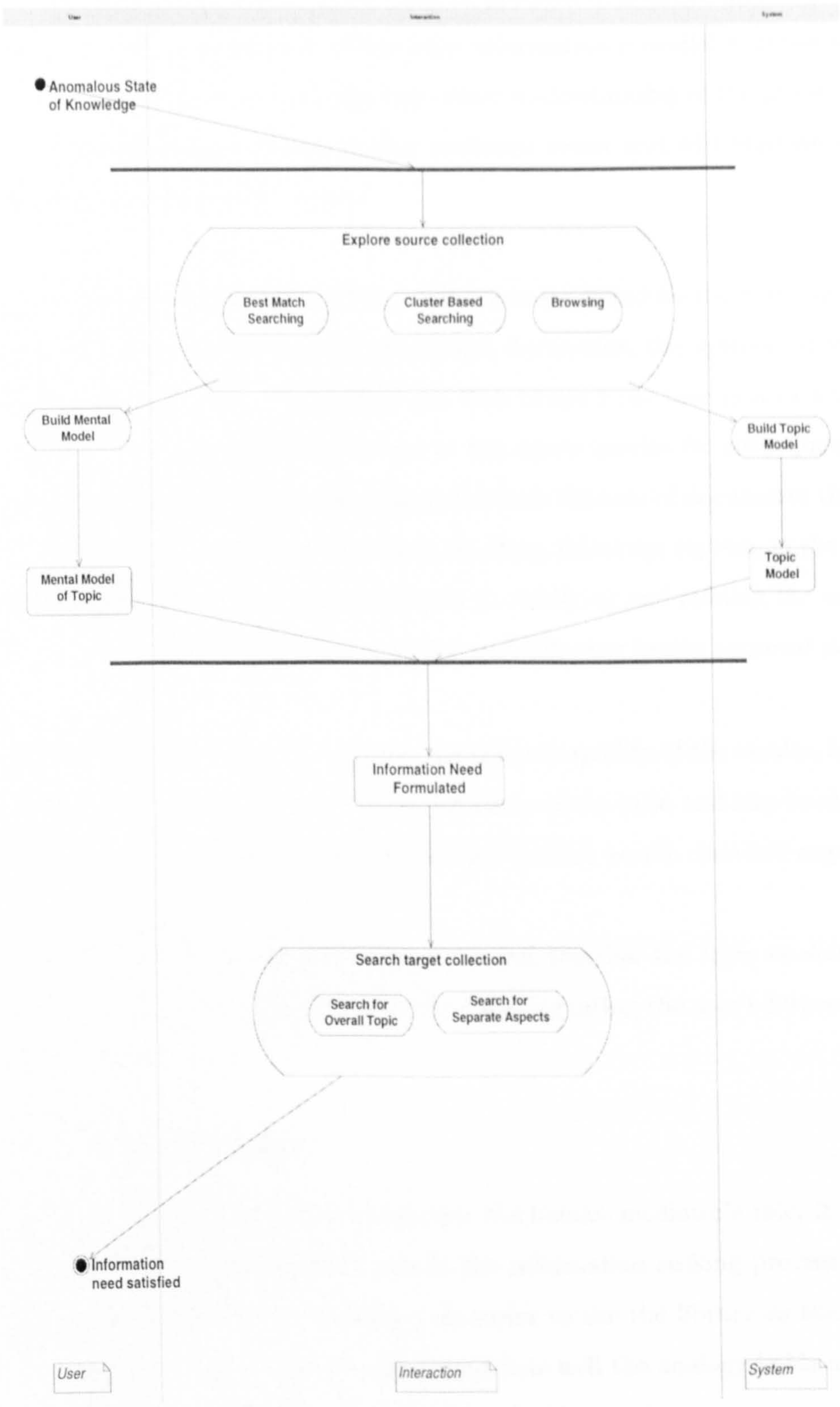


Figure 3.3: Detailed mediated process.

The user's complex interaction with the system has two consequences. Firstly, the user builds a better *mental model* of her topic or problem of interest, and may even solve the problem, completely or partially, if the right information is available in the source collection. Perhaps even more importantly, this better understanding of the problem domain will make the formulation of future similar problems easier and will improve the use of search strategies in future explorations.

Secondly, the interaction has an effect on the system: based on the user's actions, and especially on her selection or marking of relevant documents, the system builds a *model* of the user's topic of interest. The system can then support the user in a variety of ways: suggest queries or at least additional terms to the user's queries for subsequent searches on target collections, suggest the exploration of certain clusters of documents that may be relevant, or suggest various search strategies. In effect, this stage represents the *mediation stage*: the system replaces the human mediator in clarifying and refining the user's information need and in formulating better queries and following better retrieval strategies.

The mediation has potential to improve not only the quality of the results, but also the user's satisfaction in terms of perceived completeness of the task, and also to make the retrieval process more enjoyable by reducing the user's effort, search time and cognitive load.

If used over a period of time, the system can put together the *topic models* for a user and build *user profiles*, which can be used for disambiguating the user's future queries, or for monitoring tasks.

3.2.2 The library analogy

As our mediation system attempts to emulate the human mediator's role, it is useful to look in more detail at the librarian's role in the information seeking process and at the interaction that takes place in the library. In order to use the library as the *interaction metaphor* of our system, we need to understand how well the analogy holds and where it may break, in order to avoid potential confusion for the user[1].

[1]The library metaphor has been successfully used in projects such as the BookHouse [RPG94].

The typical library, as a collection of information resources, is a good metaphor for the source collection: it is of a reasonable size, it has a precise structure and this structure can be browsed in its entirety. Studies of interaction in a public library [Nor96] have identified three phases of the interaction between user and librarian:

1. problem presentation and clarification - the user formulates an initial request and responds to clarification questions from the librarian.

2. catalog consultation (by the librarian) - clarification and refinement of the request may continue here.

3. problem solution by browsing the shelf.

If the user's information need is too vague or her interest is too general to support the formulation of a query, the librarian can skip step 2 and take the user straight to the shelf with books that cover her topic of interest.

We extend and slightly alter the formulation of the search process. **Figure 3.4** shows the metaphor that we use. The steps in our model are as follows:

1. **Select a library.** Even for a vague information need, the user needs to do a rough analysis and to decide which of the available libraries offers more chances of holding the needed information, for example the Computer Science Department Library, the Business School Library or maybe the City Library.

2. **Consult catalog.** Once in the chosen library, the user has access to the catalog, either online or on cards. If the user is not confident about the organisation of the catalog, about the domains covered by the library, about the vocabulary of these domain, or if her information need is not clear, the librarian can help. Knowledgeable with regards to search strategies and also to the structure of the library and the indexing system, the librarian consults the catalog based on the information elicited from the user, and helps clarify and focus the user's information need.

3. **Browse the shelf.** Based either on the catalog, which offers starting points, or on familiarity with the library structure, the search continues at the shelf, with or without the librarian's assistance. While browsing the materials available, guided

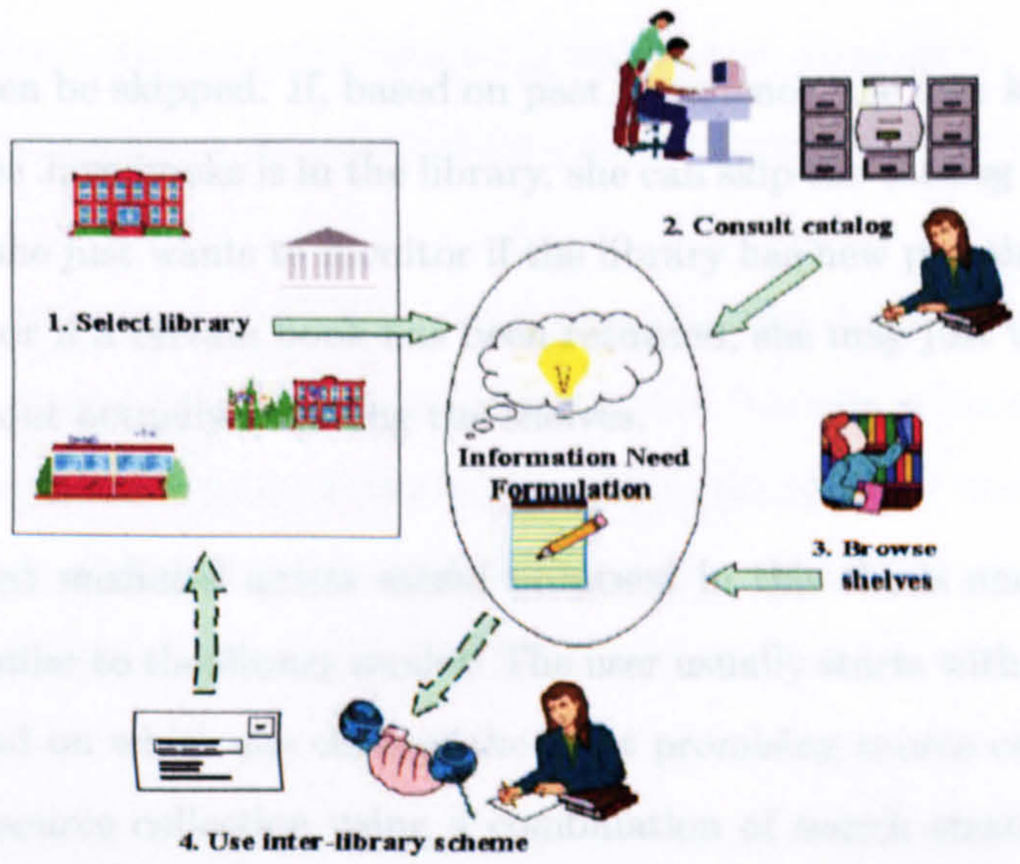


Figure 3.4: The Library metaphor.

by the physical and also topical structure of the library, the user further clarifies her information need and, hopefully, identifies the needed 'information holders'. Of course, books on shelves are grouped in topics and subtopics, so each book most probably has in its vicinity other similar books or books that treat the same or different aspects of the topic explored. There is, therefore, potential for discovering serendipitous relevant documents or information even if the starting point for browsing was not perfectly on target.

4. **Extend the search**. The library may not fully satisfy the user because some documents are temporarily unavailable, or some specific aspects of the user's topic are not covered in sufficient detail. In such a case the user can use some inter-library scheme or other information sources. What is essential is that, due to the mediation that has taken place, the context of the problem to solve is better understood and the information need is clearly formulated, so the request to other information sources is unambiguous.

There is some flexibility in the order of the steps above and in the flow of information between the steps. For example, a user browsing a shelf with Java books, looking for the implementation of a certain algorithm, may realize that a generic book on data structures and algorithms would be more appropriate, so she may go back to the catalog.

Some steps may even be skipped. If, based on past experience, the user knows exactly where the shelf with the Java books is in the library, she can skip the catalog consultation. On the other hand, if she just wants to monitor if the library has new promising books on Information Retrieval or if a certain book has been returned, she may just use the online catalog remotely, without actually browsing the shelves.

The computer-based *mediated access model* proposed in this thesis and depicted in **Figure 3.5** is quite similar to the library model. The user usually starts with an ill-defined information need, based on which she chooses the most promising source collection. She explores the selected source collection using a combination of search strategies: ranked document retrieval, cluster-based retrieval, browsing of the structure.

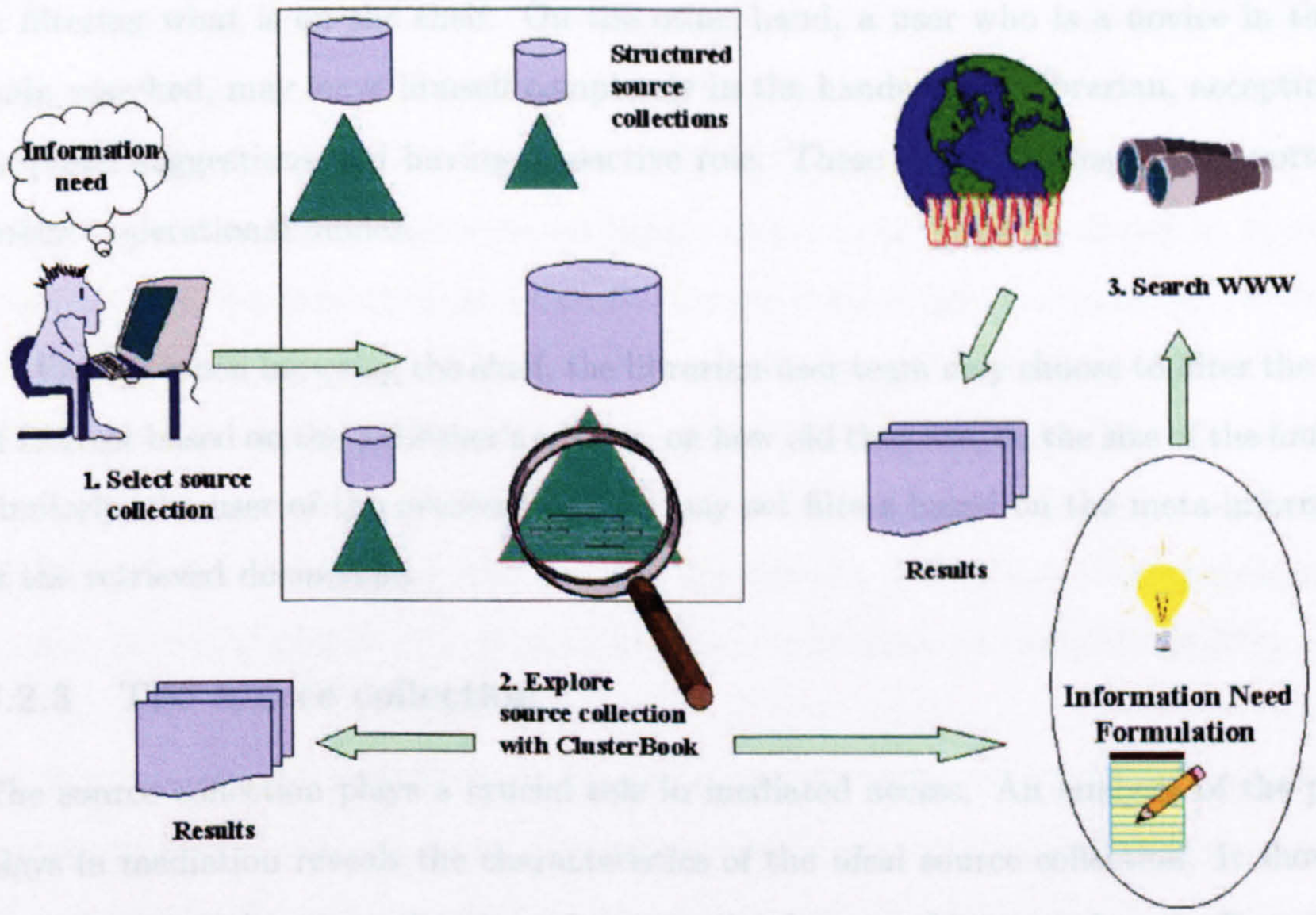


Figure 3.5: The mediation sequence.

One outcome is a set of results which may help the user solve her task. However, if the user intends to do a more comprehensive search ("What else can I find on the Web on this

subject ?") or has a monitoring task ("Has anything new been published in this area ?"), then the system comes to her assistance. Based on the *language model* of the topic investigated, built during the user-system interaction, the system generates a query or a set of queries and extends the search to the selected target collection, which is typically the Web.

Monitoring tasks are supported by bookmarking relevant documents or clusters that are representative of topics expected to be of long-term interest. These bookmarks can constitute starting points for future explorations. Topic models and even queries used on the target collection can also be bookmarked, if the user's intent or task is to periodically check the Web for new documents on a certain topic.

The interaction between the librarian and the user is also emulated by WebCluster. A user knowledgeable of the searched domain may contribute useful keywords, may accept or reject keywords or suggestions proposed by the librarian and may be very positive in filtering what is on the shelf. On the other hand, a user who is a novice in the domain searched, may leave himself completely in the hands of the librarian, accepting the proposed suggestions and having a reactive role. These two scenarios are supported by distinct operational modes.

Finally, when browsing the shelf, the librarian-user team may choose to filter the items of interest based on the publisher's edition, on how old they are, on the size of the font, etc. Similarly, the user of the retrieval system may set filters based on the meta-information of the retrieved documents.

3.2.3 The source collection

The source collection plays a crucial role in mediated access. An analysis of the part it plays in mediation reveals the characteristics of the *ideal* source collection. It should be large enough to be comprehensive relative to the domain of interest, but small enough to afford operations such as filtering, clustering, classifying or sorting in reasonable time. It should have a clear topical structure, in order to support exploration via a combination of browsing and searching, and it should be representative of the user's domain of interest, so that the user can learn the domain's terminology, its concepts and topics, and better

understand her problem and its context. In practice the ideal source collection may not always be available, so we have to examine possible choices of source collections, for various domains of user interest.

One type of source collection is the *manually classified specialised collection* that covers the user's domain of interest. There have been attempts, by libraries or specialised organisations or companies, to collect, classify and maintain collections that cover various fields. For example, MEDLINE is a collection covering Medicine, which has associated a terminology model (Unified Medical Language System - UMLS) and a hierarchical classification of medical concepts (Medical Subject Headings - MeSH)[2]. Communications of the ACM (CACM) is a collection of articles on Computing, with the associated ACM Computing Classification System[3].

Specialised collections and classifications can be seen as loosely coupled: although every specialised domain has a terminology and an underlying structure, an explicit *ontology* (or classification of the domain's concepts) may not exist. Also, how representative a specialised collection is for a universe of discourse depends on the coverage and on the depth of covering the domain's issues. Some domains are well represented by specialised collections but no ontologies are available. On the other hand, ontologies can be built for a domain by human experts based on their knowledge and experience, without a certain document collection being available for support or exemplification. The consequence is that a classification system developed for a certain domain can be used to classify any document collection covering that domain. For example, UMLS can be used to (manually or automatically) classify any medical document collection, not only MEDLINE.

If no ontology and no manually classified representative collection are available for a certain domain (or even if there are), an alternative approach to structuring a collection in order to build a source collection for mediation is *clustering*, or automatic classification. For example, Cranfield is a collection of abstracts on aerodynamics, RAPRA covers polymers, CABI covers agriculture issues and so on. These collections can be clustered in order to reveal the topical structure of their domain and to support exploration of the

[2]http://www.nlm.nih.gov/pubs/factsheets/pubmed.html
[3]http://www.acm.org/class

documents they contain. In the process of clustering, support tools like an index or a thesaurus can be built for supporting exploration of the vocabulary and the topical structure of the domain. A critical issue is choosing the clustering method and parameters, the similarity measure between documents as well as the method for creating cluster representatives that are representative, accurate and have discriminatory power. These will be addressed later in the thesis.

A *mixed* approach to building the source collection based on a document collection can also be imagined. In the first stage, a clustering algorithm is used in order to automatically group documents based on reciprocal similarities, as estimated by the system. In a second step, human experts can adjust the obtained structure by exploring the structure and moving to the right cluster the documents whose semantic content does not fit its place, or making copies for documents that should belong to more than one cluster.

The idea of *exemplary documents*, proposed by Blair and Kimbrough [BK02] is also an interesting potential methodology for producing a source collection for mediation. It addresses a similar situation to the one addressed by the WebCluster case: the user seeking information in a large target collection that only affords query-based searching, in order to satisfy a task in a certain domain. If unfamiliar with the domain, the user does not know the intellectual, topical structure of the domain and is extremely constrained in her investigation. All she can do is try to guess words that may appear in relevant documents (but do not appear in non-relevant documents) and adapt her query according to the retrieved set. Such unsystematic piecemeal process is not appropriate for learning, as the user may never be able to see enough documents to form a general opinion about how documents are represented and, more importantly, she can only see individual documents, which makes it hard to infer semantic relationships between documents and the topical structure of the domain explored.

The theory behind the exemplary documents proposal is based on Wittgenstein's theory of language acquisition, which states that language is acquired not by definitions and explanations alone, but by having the terminology and expressions in question demonstrated in ordinary or typical use. The authors of the cited paper discuss possible kinds of

exemplary documents that can offer an intellectual road-map to the semantic content of a domain: survey articles, editorials, opinion papers, lead articles, seminal papers. However, they do not propose a (manual or automatic) procedure for reliably identifying these documents, or for building a 'domain model' that could subsequently support retrieval of other relevant documents in the domain. They also leave un-answered the problems of guaranteeing comprehensive *coverage* of the domain, or of covering the domain topics at various levels of granularity. These shortcomings can be seen as open questions to an interesting approach.

It must be said that WebCluster's approach of using a relatively small, specialised collection as the source collection for mediation has exactly the same aim of guiding the exploration of a domain of interest. We envisage the use of specialised document collections as potential exemplary documents and support the user's exploration by structuring these collections.

If no explicit source collection is available, one can be built from the target collection. A possible approach is to produce a sample of the target collection that is representative enough and covers all the sub-domains, topics and concepts of the domain, but at the same time is small enough to afford (manual or automatic) classification and exploration through a combination of searching and browsing. Another approach is to apply an initial user query as a filter and to classify the obtained source collection on the fly. Consideration must be given to the fact that the user may not have a clear information need or a good grasp of the domain vocabulary. Therefore, the filtering should be rather 'generous', including in the retrieved set even documents with a low estimated relevance for the initial query. The user should also be encouraged to supply as many words as possible, or a thesaurus should be used for query expansion.

Other issues regarding the source collection, addressed in the WebCluster project, are outwith the scope of this thesis, so they will be just mentioned for the sake of a complete image over the mediation process. One is the issue of the user selecting the appropriate source collection from the ones available. The challenge is to communicate to the user the domains covered by the source collections and to recommend the ones

that best match the user's topics of interest. Another issue is the ownership of the source collection. For public collections, the content and structure of the source collection (built by a system administrator) should be suitable for the general user, and the user should not be allowed to permanently modify it. If the user owns the collection, then more flexibility is allowed. The content can be modified, or even built from scratch, for example, by the system automatically saving documents retrieved and opened by the user. The user can manually classify a personal collection, or can use automatic classification, or clustering, and obtain a structure appropriate for exploring her domains and topics of interest.

3.2.4 Structuring the source collection

There is quite a rich literature on interactive information retrieval systems based on manually classified collections (mainly MEDLINE [HLH94, Pra99, PHF99]). On the other hand, document clustering has been studied mostly in the context of batch retrieval [Wil88] or as a tool for structuring search results [HP96, ZEMK97, ZE99]. There is little understanding of the power of clustering to reveal the topical structure of a document collection and to guide exploration in an interactive setting. To fill this gap, this thesis will focus on the use of document clustering in an interactive information retrieval system, for exploratory tasks, and will explore the potential of document clustering for structuring source collections for mediated retrieval.

The expected usefulness of document clustering is based on the **cluster hypothesis**, i.e. on the expectation that "closely associated documents tend to be relevant to the same requests". Most researchers whose work was based on this hypothesis and who tried to evaluate its validity assumed a reciprocal (bi-directional) relationship between similarity and relevance, i.e. similar documents are expected to be relevant to the same queries and documents relevant to the same queries are expected to be similar. This assumption is implicit in the overlap test proposed by the authors of the hypothesis[4] and is made explicit by El-Hamduchi and Willett [EHW87] ("dissimilar documents are unlikely to be relevant to the same requests") and Hearst and Pedersen [HP96] ("relevant documents tend to me more similar to each other than to non-relevant documents").

[4]See the review chapter.

Consequently, there has been no or little distinction between experiments attempting to show that similar documents tend to be relevant to the same topics and experiments testing whether documents similar to the same topics are highly similar. Moreover, most experiments on cluster-based retrieval have looked at the distribution of topical documents over the cluster structure, with no or little distinction between the capacity of the clustering algorithm to group similar documents, and the relationship between similarity and topicality. There is little surprise, therefore, that such experiments have produced inconsistent results.

Informal experiments with WebCluster suggest a uni-directional relationship and a *relaxed cluster hypothesis*: similar documents tend to be relevant to the same requests, but documents relevant to the same requests are not necessary similar. They tend to be dissimilar if they cover different *aspects* of the same complex topic.

Our experiments also paint a slightly different picture of document retrieval compared to the one built by the literature review. Let us call *features* the sets of terms (or keywords) representative for a certain *topic* (or *aspect* of a topic, for complex topics). Documents are represented by features, the inter-document similarity is computed based on features and the clustering is generated based on features. Based on their contribution (or weight) one can distinguish between major features, which determine the major axes or higher level clusters of a hierarchical structure, and minor features, which determine the minor axes or bottom level clusters of the hierarchy. For example, clustering a sub-collection of Reuters produces two major clusters of documents that refer to the former US president Reagan: one about the Iranian arms deal, and one about the American agriculture. In both cases, "president Reagan" is just a minor feature.

A query that matches a major feature ("Iranian arms deal") is very likely to hit a major cluster, so that most relevant documents are grouped together, so cluster-based retrieval gives good results. On the other hand, if the query matches a minor feature ("president Reagan"), then the relevant documents are indeed grouped into small subclusters, but spread over the collection. The documents about president Reagan in the "Iranian arms deal" cluster are quite similar to each other, but dissimilar to the ones in the "American

agriculture" cluster. In this case, classic cluster based retrieval, which retrieves just one cluster, would do badly. A relaxed version of cluster-based retrieval, more appropriate in an interactive context, would rank the clusters of the structure based on their estimated quality, and allow the user to explore them.

Our relaxed cluster hypothesis is in agreement with the results of traditional experiments on the original cluster hypothesis. Early experiments used small test collections and very simple and focused requests. In such cases, with topics that only had one or a small number of aspects, the cluster hypothesis experiments were successful. Later experiments, with larger collections and more complex topics, were less successful; this is what our hypothesis would predict.

This thesis also conjectures that it is possible to alter the clustering axes by applying an external weighting scheme. For example, by increasing the weight of terms extracted from *frequently asked questions* it is expected that the 'hot topics' will become more visible in the structure, i.e. the documents relevant to those topics will be better grouped together. An experiment is needed to evaluate this hypothesis.

Note that different weighting schemes for clustering or different classification schemes can be used for mediation systems targeted at the general public, respectively at specialised groups. For example, a collection of medical articles can be clustered differently for the interest of various groups, making prevention methods the main axes of the structure for some users and treatment methods for others.

WebCluster takes into account this new view of clustering, by allowing the user to collect (or berrypick [Bat89]) and bookmark subclusters of interest from 'pockets of relevance'. We will call this *aspectual retrieval*, as it usually identifies aspects of a topic or of a request.

Figure 3.6, detailing the first stage of mediated access i.e. the exploration of the source collection, illustrates this issue: relevance is often *aspectual*, in the sense that a document can be relevant to a query for a variety of reasons, potentially responding to

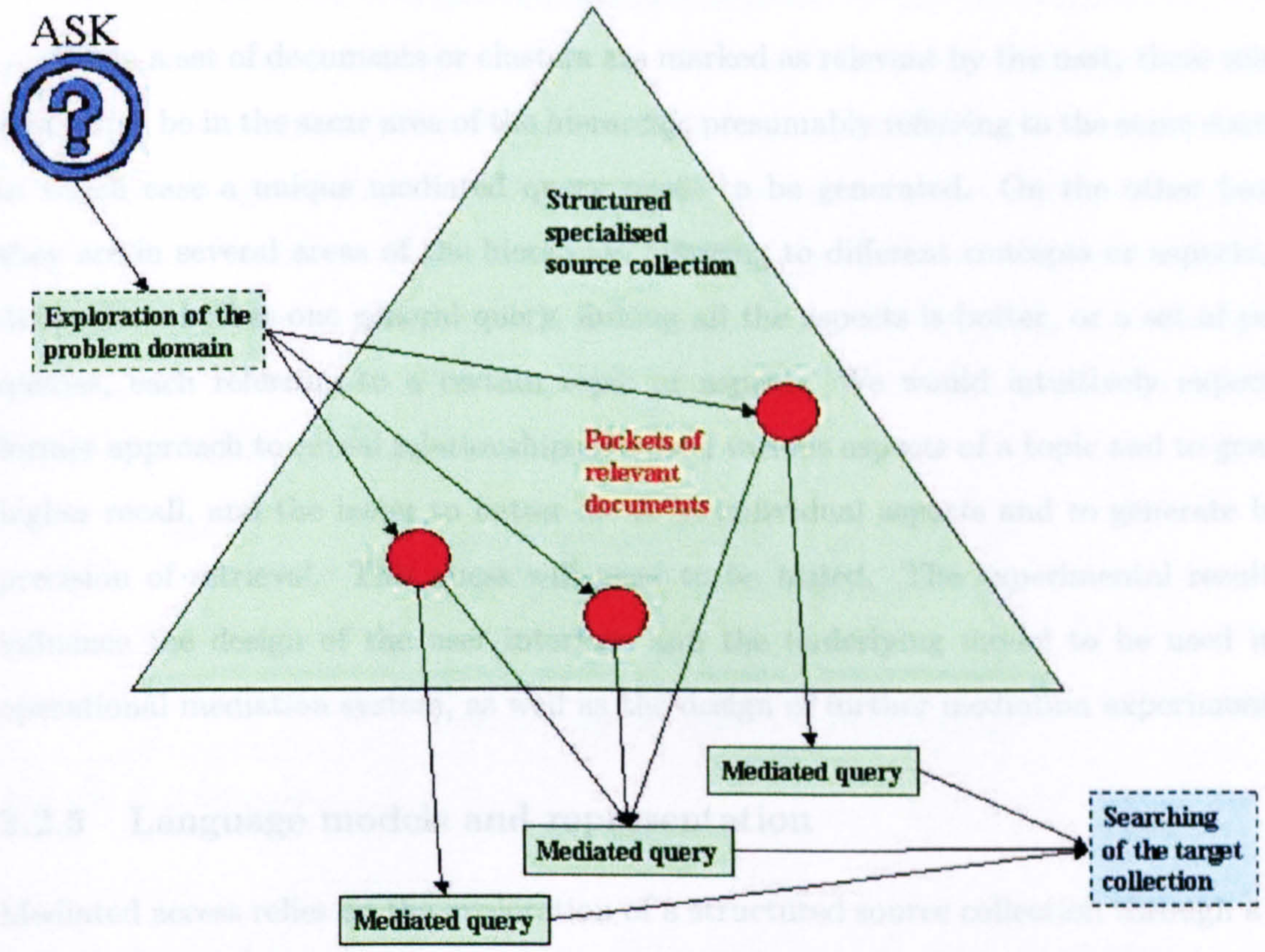


Figure 3.6: Aspects of relevance in the mediated access process

various aspects of an information need. For example, documents about ethnic violence in South-East Asia, documents about the Indian Ocean climate and documents about the Passport Agency's problems, although very dissimilar, may all be very relevant for someone planning their holiday. Although similar documents can be expected to be relevant for the same queries, the reciprocal statement may be not true: documents relevant to the same query need not be similar. Intuitively, they are similar if they refer to the same aspect of the information need and dissimilar otherwise.

This issue applies to the generation of the mediated query and also to the design of the user interface. Imagine the user who wants to go on holiday. She may not be aware of problems at the Passport Agency, of safety issues (either because of violence or because of some disease outbreak), of the monsoon period, or of some cheap deals. However, a general query such as "holiday Asia" should point the user to all these various aspects that should be of interest to her, so that she explores all potentially useful pockets of

information in the source collection.

When a set of documents or clusters are marked as relevant by the user, these selected items may be in the same area of the hierarchy, presumably referring to the same concepts, in which case a unique mediated query needs to be generated. On the other hand, if they are in several areas of the hierarchy, referring to different concepts or aspects, it is debatable whether one general query, linking all the aspects is better, or a set of precise queries, each referring to a certain topic or aspect. We would intuitively expect the former approach to reveal relationships between various aspects of a topic and to generate higher recall, and the latter to better focus on individual aspects and to generate better precision of retrieval. This guess will need to be tested. The experimental result will influence the design of the user interface and the underlying model to be used in the operational mediation system, as well as the design of further mediation experiments.

3.2.5 Language models and representation

Mediated access relies on the exploration of a structured source collection through a combination of searching and browsing, and on the automatic generation of the *mediated query*, based on the documents and clusters marked as relevant by the user. Searching and browsing rely on document and cluster labels being representative for their content; the mediated query also has to be representative for the set of relevant documents.

Therefore, an essential issue is the representation of the content for documents, clusters and collections. Documents and collections can be viewed as particular cases of clusters (having a single document and, respectively, all the documents) so it is sufficient to discuss the generation of *cluster representatives* (also called *labels* or *centroids*).

For all IR systems using clustering, as found in the literature, a unique representative was considered for each cluster, containing terms deemed typical for the cluster, usually based on their frequency. Both un-weighted [JR71] and weighted [Voo85b] centroids have been described. However, getting the right balance between *accuracy* in representation and *power of discrimination* is problematic and can involve adjusting thresholds and regenerating representatives for particular applications.

Our novel approach is to generate multiple representatives, each adapted to a specific purpose: browsing, searching or mediation. Due to their power, flexibility, and uniform treatment of document and clusters, statistical language models[5] are a natural choice as a technique for generating representatives. The formulae that we employed for the label generation are based on the *Kullback-Liebler (KL) divergence* or *relative entropy*, which indicates how different two probability distributions are [MS99, p.72]. In our context, if P and Q are clusters, viewed as bags of terms, then for each term t_i we can calculate its probability distribution in the two bags:

$$p_{i,P} = \frac{number\ of\ occurences\ of\ term\ t_i\ in\ P}{total\ number\ of\ term\ occurences\ in\ P},$$

$$p_{i,Q} = \frac{number\ of\ occurences\ of\ term\ t_i\ in\ Q}{total\ number\ of\ term\ occurences\ in\ Q},$$

and the KL formula

$$KL_i = p_{i,P} \log \frac{p_{i,P}}{p_{i,Q}} \tag{3.1}$$

indicates the *relative specificity* of t_i in P, compared to Q. The terms that have positive values for this measure are more specific to P than to Q and high levels of relative specificity indicate terms that are much more typical for P than for Q. The set of terms weighted and ranked according to KL form the representative of P in the context of comparing P with Q.

In the rest of this subsection we will explore how the Kullback-Liebler divergence can be used to generate various cluster representatives.

Relative cluster representative for browsing

Imagine a user browsing the hierarchic cluster structure. In order to decide which of the subclusters of the current cluster is worth expanding for further exploration, she needs to know what is specific about each subcluster. For that she relies on the cluster labels displayed in the user interface. Therefore, the browsing label of each cluster needs to indicate in what way the cluster differs from its parent. This suggests the use of the

[5]Language models were reviewed in section 2.2.2.

Kullback-Liebler divergence measure between the probability distribution in the cluster, and the corresponding probability distribution in the parent, for each term t_i in the cluster:

$$R_i = KL_i(cluster, parent) = p_{i,cluster} \log \frac{p_{i,cluster}}{p_{i,parent}}. \tag{3.2}$$

This weight indicates the relative specificity of each term in the cluster, compared to the parent cluster. The terms with negative weight are ignored (they are not specific) and the remaining terms are ranked according to their R_i weights in order to generate the browsing label, or *relative representative.*

Absolute cluster representative for searching

When searching the source collection, based on the user's query, the system needs to find the cluster that best matches the query, and therefore the representative needs to distinguish each cluster from the rest of the collection. We therefore compute the term weights of the *search label* by applying the KL formula between the term probability distribution in the cluster, respectively in the collection:

$$A_i = KL_i(cluster, collection) = p_{i,cluster} \log \frac{p_{i,cluster}}{p_{i,collection}}. \tag{3.3}$$

The terms with negative weight are ignored and the remaining terms are ranked according to their A_i weights in order to generate the *absolute representative.* Conceptually a cluster-based search is performed by matching the query with each of these search representatives or labels. While such an approach is appropriate for a top-down search strategy, an implementation of a comprehensive search strategy would probably be more efficient if an inverted file of cluster labels was built.

Expanded cluster representative for mediation

The absolute labels appear to be adequate when conducting cluster-based searching on a collection. Imagine, however, that a user has simultaneous access to a set of distributed document collections, covering various domains, and that she submits a common query to all of them. For example, a user interested in applications of graph-matching algorithms to molecule matching may choose to simultaneously search specialised collections covering mathematics, computer science and biochemistry, as well as the intranet of her company

covering all of these. The search algorithm returns a list of clusters, ranked according to their estimated relevance to the user's query. In order for the user to decide which clusters look promising and are worth investigating in detail, the cluster representatives need to convey the content of the clusters, but also their context.

Our approach is to employ a *combined model*, by summing gradually reduced contributions of the absolute representative of the chosen cluster, of its parent, and of all the clusters on the path to the root of the collection's hierarchic structure. The weight of term t_i in the *expanded representative* is:

$$E_i = (1-w)\cdot A_{i,0} + (1-w)\cdot w\cdot A_{i,1} + (1-w)\cdot w^2\cdot A_{i,2} + \ldots$$
$$+(1-w)\cdot w^{r-1}\cdot A_{i,r-1} + w^r\cdot A_{i,r},$$

where $A_{i,0}$, $A_{i,1}$, ..., $A_{i,r}$ are the weights of t_i in the absolute representative of the chosen cluster, its parent, ..., the root cluster, and $w \in (0,1]$ is the decay rate of the contribution as the context goes from specific to general. For example, for $w = 0.1$, the contribution of the current cluster to the term weights is 0.9, of its parent 0.09 and so on.

When applying the combined model, all the terms in the vocabulary are considered, not only the terms in the selected cluster.

Document representatives

As mentioned above, documents in the clustered source collection can be treated as clusters with just one document for the purpose of building representatives and the above formulae can be applied.

However, the Kullback-Liebler divergence measure (or relative entropy) can also be used as an alternative to the classical *td-idf-dl* family of formulae in contexts that have nothing to do with clustering. In WebCluster this is the case with the target collection, which is too large to be clustered, but needs to be indexed in view of searching.

Searching a collection C of documents d based on a query q means ranking the docu-

ments based on their estimated relevance to the query. This is done by computing a score for each document, according to how specific the query terms are for the document.

The relative frequency of a term t_i in a document d, which is the probability of the term being generated by a random generation process based on the document,

$$p_{i,d} = \frac{f(d, t_i)}{|d|},$$

indicates the specificity of the term in the document. The relative frequency of a term t_i in the collection C , which is the probability of the term being generated by a random generation process based on the collection:

$$p_{i,C} = \frac{number\ of\ occurrences\ of\ t_i\ in\ C}{total\ number\ of\ term\ occurences\ in\ C},$$

indicates the specificity of the term in the collection. The Kullback-Liebler divergence measure between $p_{i,d}$ and $p_{i,C}$ indicates whether the term t_i is more specific in the document d or in the whole collection C:

$$W_{i,d} = p_{i,d} \log \frac{p_{i,d}}{p_{i,C}}. \tag{3.4}$$

For example, if the word "computer" appears very often in an isolated document, that term may be considered to be very specific for the document. However, if the document is part of a collection on Computer Science, and the term "computer" appears throughout the collection, it is not a good indexing term for any document, as it does not help to distinguish documents from each other. This approach naturally suggests a weight threshold in building document representatives: only the terms t_i for which $W_{i,d} > 0$ should be used for representing document d, and the terms should be ranked in the decreasing order of the weight.

Therefore, the KL formula appears to be quite appropriate for computing documents term weights and, implicitly, document representatives. Consequently, the scores allocated to documents, in the search process, are based on the weights of the query terms in the

documents:

$$KL(d, C, q) = \sum_{f(q,t_i)\neq 0} KL(d, C, t_i),$$

where

$$KL(d, C, t_i) = \sum_{f(d,t_i)\neq 0} \frac{f(d, t_i)}{|d|} \log \frac{f(d, t_i)/|d|}{f(C, t_i)/|C|},$$

with $f(d, t_i)$, $f(C, t_i)$ and $f(q, t_i)$ being the frequency of term t_i in the document d, the collection C and query q.

When implementing a search procedure based on the formulae above there is a choice between computing the contribution $KL(d, C, t_i)$ of each query term to the document scores at search time, when a query is submitted, and computing document representatives, which contain the weight of each term in the vocabulary in each document, at indexing time. Sometimes building the explicit document representatives is necessary, for example when computing the similarity between documents for nearest-neighbour searching, or in view of subsequent clustering.

We will evaluate the effect that a weighting scheme based on language models has on the quality of the clustering, in comparison to one based on traditional *tf-idf-dl* schemes.

3.2.6 Topic models

Topic representation

The conceptual model of mediation access relies on the user making relevance judgements with regards to documents and clusters of the source collection, and thus conveying her topic of interest. In effect, the user provides an *exemplary* representation of the topic of interest, which consists of documents and clusters of documents that are typical for the topic investigated. The system performs a statistical analysis of the exemplary documents and derives a *statistical* or *language model* representation of the topic. This consists of the probability distributions of the terms in the vocabulary over the 'typical' topical document. Based on the context (the source collection and the target collection, as well as the assumptions made for the user interaction), the system derives a *keyword* representation of the topic, which consists of the terms that discriminate the topic in the given context,

ranked based on their weight, or contribution to the topic. Finally, the mediated query, used for searching the target collection, is derived from this keyword representation of the topic, typically by applying a threshold on query size or term weight[6].

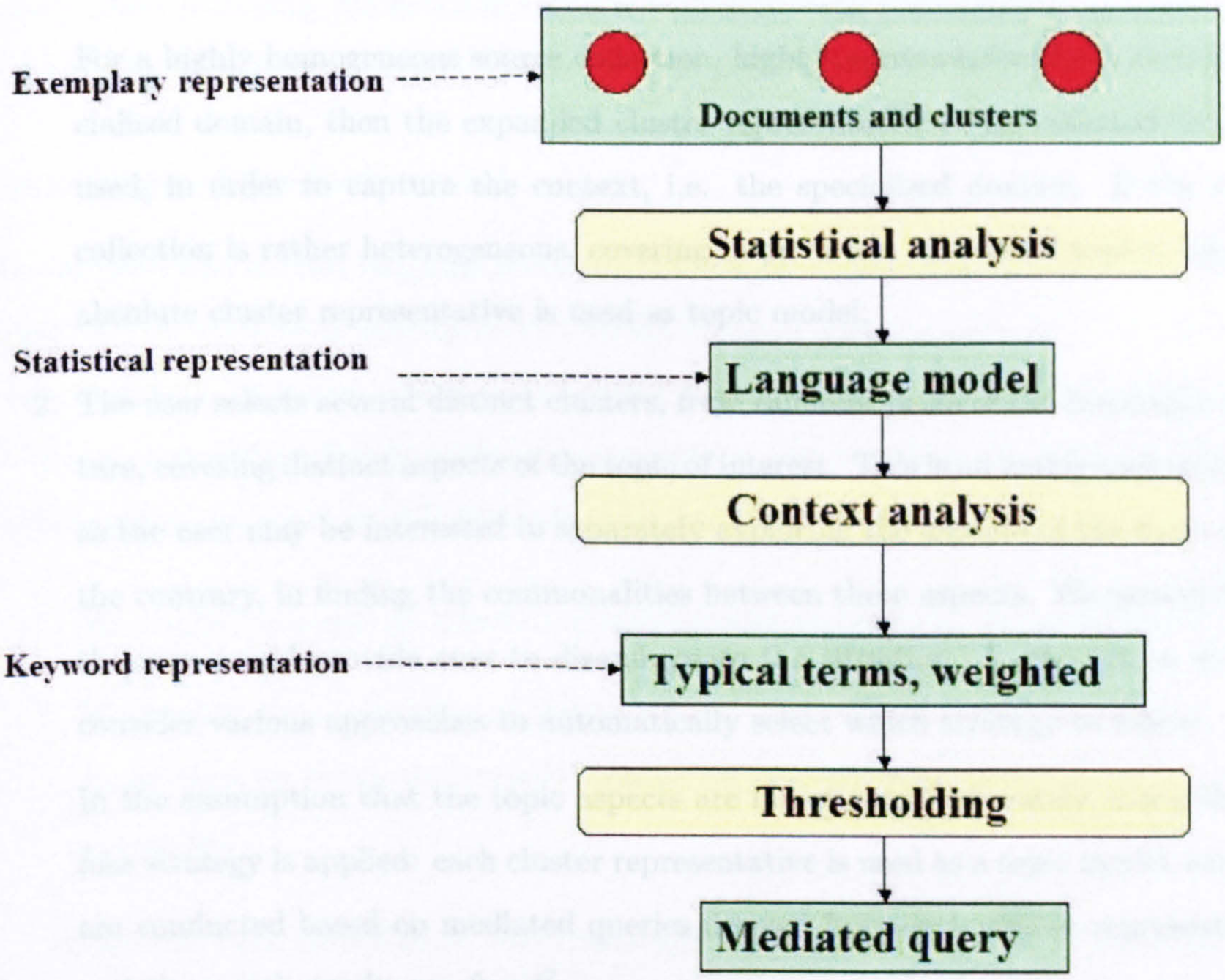


Figure 3.7: Topic model representations

This model, depicted in **Figure 3.7**, is expected to be representative for the mediated interaction. Shortcuts are, however, acceptable: as the user explores the source collection and learns the terminology and concepts of the problem domain, she may become confident enough to generate the mediated query herself, rather than rely on the system completely.

The generic model also leaves room for variation, especially with regards to taking the

[6]The size of the query can be determined by a trade-off between recall and precision: while highly-topical terms are expected to return high precision, lower weight terms are likely to increase recall, but with a potential decrease in precision.

context into account. For the proof-of-concept system designed as part of the WebCluster project, we considered two cases:

1. The user selects just one document or cluster. In this case the document or cluster representative is used to represent the topic. The exact formula that we advocate depends on the source collection homogeneity and specialisation.

   For a highly homogeneous source collection, highly representative for a certain specialised domain, then the expanded cluster representative of the selected cluster is used, in order to capture the context, i.e. the specialised domain. If the source collection is rather heterogeneous, covering more or less un-related topics, then the absolute cluster representative is used as topic model.

2. The user selects several distinct clusters, from different parts of the hierarchic structure, covering distinct aspects of the topic of interest. This is an ambiguous situation, as the user may be interested in separately exploring the aspects of the topic or, on the contrary, in finding the commonalities between these aspects. We assumed that the user would provide cues to disambiguate the situation. In the future we may consider various approaches to automatically select which strategy to follow.

   In the assumption that the topic aspects are investigated separately, a search-and-fuse strategy is applied: each cluster representative is used as a topic model, searches are conducted based on mediated queries derived from each cluster representative, and the search results are fused[7].

   If the commonality between aspects is to be investigated, the fuse-and-search approach is needed: a virtual 'relevant cluster' is built from the documents selected as relevant, its cluster representative is computed and taken as representative of the topic, and the derived mediated query is used to search the target collection.

In this discussion we have assumed the WebCluster context, in which the user explores a clustered source collection. The idea of building topic models and mediating the user's search can be extended to other scenarios, or other ways for the user to explore the source collection and select relevant documents. If the source collection has been manually classified or structured in another way, then the categories can be used instead of the clusters.

[7]Fusing the results is a research problem on its own, so we are not addressing it here. In our implementation we used simple score-based fusion.

The categories may have been assigned labels during the classification process, which can be used for mediation, or alternatively the topic models can be generated based on the documents contained in the category.

Even if the source collection is not structured in any way, but can be explored (by query-based searching, for example), then the set of exemplary documents selected by the user can be viewed as a relevant cluster and our procedures can be applied.

Alterations of the basic model

Researchers who applied language modelling techniques to Information Retrieval have also tried various *smoothing techniques* to account for data sparsity as well as for synonymy. For example, a document about motor-racing may not contain either the term "fast" or "vehicle" and still be highly relevant for queries on fast vehicles. We have used no such techniques in our topic models and this is for two reasons:

1. We wanted to avoid an inflation of independent variables in our experiments.
2. We wanted to concentrate on validating the basic concept and model rather than spend time on trying out improvements that could distract from the main focus of the thesis.

While the work described in this thesis has obvious limitations, it is envisaged that future work will attempt to expand our model in order to increase its accuracy and, consequently, the retrieval effectiveness.

We did consider one alteration to the basic model, by taking into account the uniformity of the term distribution in a cluster. For example, consider a cluster with 10 documents. A term t_1 may appear once in every document, while another term t_2 may appear 10 times in one document and not at all in the other documents. While both terms have the same frequency, t_1 is more uniformly spread, which may indicate that it is more typical for the cluster.

While the basic model, like most language models that we have found in literature, views collections, clusters and other such groups of documents as bags of terms, our

alteration consists in employing a uniformity factor:

$$u = \frac{1}{1 + k \cdot \sigma},$$

where σ is the standard deviation of the term frequency over the documents of the cluster and $k \geq 0$ a parameter that can be set to indicate how important uniformity is for specificity (if $k = 0$, then $u = 1$, so there is no influence). This multiplicative factor can be used to alter the weights of terms in the topic model.

3.2.7 Query-by-example vs. explicit query formulation

As we have shown in a previous section, there is controversy between contradictory results with regards to the level of control users want during an interactive retrieval session. Some studies have shown that users prefer to have control during interaction with an IR system [Bat90, KB96] and to understand at least some of the internals of the retrieval process, such as the query expansion generated by relevance feedback. Other researchers consider that users are task-oriented [MDKL93] and want 'magic' [Cro95] rather than an understanding of how the system retrieves the results.

We will support both cases, by implementing an *opaque* as well as a *transparent* mode of operation. This refers to the visibility of the keywords representation of the topic:

- In the opaque (*query-by-example*) mode, the user does not see the topic model. After indicating the documents that she considers relevant, she asks for "more like this". The system automatically derives a *mediated query* from the topic model generated and submits it to the target collection's search engine.

- In the transparent mode, the user selects terms from the set proposed by the system (the top ranking terms in the keyword representation), builds a query, and submits it to the search engine.

The transparent mode is expected to be favoured by users who understand the search process, have a familiarity with the terminology of the domain and want control over what aspects of the topic the search should focus on. On the other hand, the opaque mode is expected to attract more novice searchers, who are happy to ignore what happens behind

the scenes, to select some exemplary documents and to get more similar documents back.

We argue that, paradoxically, the more expert users that want control and chose the transparent mode may get less from the system in terms of search accuracy. Various studies have shown that the query development is the most critical factor in retrieval [SJ00] and that improving the query weights can significantly retrieval effectiveness [Har92, BYRN99]. While in opaque mode the query is weighted based on the term weights from the topic model, the user's freedom in the transparent mode creates a more delicate situation. If the user is allowed to mix-and-match some of the terms proposed by the system and her own terms, it is impossible for the system to weight the query in a sensible way. One approach would be to restrict the user to rejecting terms proposed by the system, and not add her own; in that case the system can use the weights from the topic model.

3.2.8 Exploration strategies

When designing a mediated access system we intend to emulate the user's interaction with a human mediator. This interaction, despite its generic steps, can vary widely according to the specific task as well as to the user's personality and domain knowledge.

When designing a mediation system we should not prescribe one certain strategy, but give the user some freedom. Based on the analysis of various search scenarios considered, we have identified two distinct exploration strategies that need to be supported:

fuse and search - The user explores the source collection, trying to identify all the aspects that seem relevant to her topic, and marking along the way all the relevant documents or clusters of documents. When she is satisfied that most aspects of her topic of interest have been covered, the user asks the system for assistance in building a topic model that conveys the common theme of the marked documents and in generating a query that expresses her interest. Intuitively, we expect this strategy to provide a 'recall device'.

search and fuse - The difference from the previous strategy is that the user wants to distinguish between the various aspects of her information need. Therefore, she explores these aspects separately, marking documents relevant for each of them and

generating a distinct mediated query for each. The individual sets of results can then be fused, in order to provide coverage of all the aspects of interest. We expect this strategy to be appropriate when the user is only interested in some aspects of a topic, but wants to investigate each aspect in detail. Therefore, we expect this strategy to be a 'precision device'.

3.2.9 A discussion of the mediated access paradigm

The design decisions taken for implementing our proof-of-concept WebCluster system may wrongly convey the idea that these decisions are inherent constraints of mediation and may create an incorrect, narrow view of the mediation paradigm. Hereby we attempt to clarify this issue.

The core of the mediated access model is simple: the user explores a structured, specialised source collection indicating, in the process, exemplary relevant documents. In response, the system analyses these documents and proposes a mediated query, which the user can use to search a target collection. Based on this simple model, a variety of semi-independent choices can be made with regards to different aspects of mediation:

Structuring the source collection. We chose document clustering because it is fully automatic and domain independent, it allows variability in choosing parameters such as the indexing method, the weighting scheme and the clustering algorithm, plus it gave us the opportunity to conduct experiments on the cluster hypothesis. However, any classification method, manual, semi-automatic (supervised) or fully automatic (un-supervised), can be employed as long as it reveals the topical structure of the source collection and thus supports the exploration of the problem domain.

Exploring the structured source collection. We chose the folder metaphor for representing the structured source collection in the user interface because of its simple implementation and the computer users' familiarity with it. We also combined this structured view of the domain with a linear view, obtained through query-based ranked searching, in order to offer a rich set of retrieval strategies. However, alternative visualization tools such as hyperbolic trees, thematic maps, tree maps or cone trees can be used instead and support for other search strategies can be offered.

The constraint is that the user interface should support concept learning and the identification of exemplary relevant documents.

Explicit vs. implicit relevance feedback. We chose explicit relevance feedback because, despite the user interface complexity that it introduces and the need for user cooperation, it indicates the user's preferences more clearly. Implicit feedback, which relies on cues from the user's behaviour and actions, are currently unreliable in conveying the user's interest. However, when technology that supports it matures, this approach will also be applicable to mediation.

The generation of the mediated query. Language models were our choice due to their power, flexibility and uniform treatment of documents and clusters of documents. Alternatively, any techniques developed for query expansion based on relevance feedback can be used just as well.

The interaction model and metaphor. The traditional library was used as model and metaphor for our implementation of mediated access because of people's familiarity with it and also because our system can be seen as a replacement for the human mediator in the library. For the new generation of information seekers, arguably more familiar with digital libraries and hypermedia than with the traditional library, a different metaphor may be more appropriate. For example, an *electronic encyclopedia* could support a wide variety of information retrieval strategies and the mediated query could be used to expand the quest for information to the Web.

This discussion, and the description of the mediation model's core, reveal the clear distinction between the novel paradigm that we propose, and relevance feedback. RF is a technique for iteratively improving the query and, implicitly, the retrieved set of documents during an interactive search session on a target collection. It relies on the user having a minimum knowledge of the domain investigated and on her being able to generate an initial decent query.

In contrast, mediated access proposes a new interaction model. It relies on the user interacting with a structured specialised 'source collection', representative for the user's problem domain and rich in documents that can support the user's learning of the terminology, concepts and topics of the domain. It also relies on the user finding sufficient

exemplary documents to solve her problem or to clearly convey her information need so that a high-quality mediated query can be generated for searching the target collection.

3.2.10 Applications of mediated retrieval

Some criticism generated by seminars on mediated access and WebCluster relates to the approach of proposing a new technology and then trying to find applications for it. Some 'user-centred' researchers would have preferred an analysis of the user requirements for a certain type of users performing certain tasks, before proposing mediated access as a solution. We have a different opinion. A detailed user and task analysis is necessary when applying existing technology to solve a specific problem, in a particular operational setting. Our purpose is different. There is plenty of evidence, especially based on analysis of search engine logs [JAS98, JSS00], indicating a rather general inability of users to formulate good queries and suggesting the need of tools for supporting information exploration for a wide range of user types. Mediated access through a structured collection is proposed as a generic solution, and its potential is investigated for various types of generic tasks and users. Based on the investigation conducted as part of this thesis, it will be possible to establish guidelines to indicate the appropriateness of employing the proposed tool in concrete situations. More in-depth experiments will also indicate optimizations of parameters for the tool and also details of particular implementations.

There is potential for a variety of applications based on the concept of mediated retrieval. A main one is *specialised content portals* to the Web. Various *content providers* put together collections of documents, abstracts or bibliographic references in domains as diverse as agriculture, finance, health care, engineering, psychology and make money by offering access to these collection. Using the WebCluster approach, these specialised collections can support the exploration of specialised domains and can extend specialised searches to the Web. The user would select documents on a topic of interest and would ask for "more like this" from the Web. The system would formulate a precise query and would send it to a Web search engine. What happens, in effect, is that the specialised source collection acts as a filter on the Web and it also imposes the structure of the specialised domain to the Web documents. The tasks supported range from exploration of the domain by novice users to monitoring of certain topics on the Web by experts who

have bookmarks in the source collection.

These specialised content portals rely on the existence of specialised collections and on visualization tools to explore them. If these collections are already classified manually, they are ready to be used as source collections, otherwise clustering needs to be applied in order to reveal the semantic, topical structure of the domain represented by the collection. A system administrator would apply various clustering algorithms offline, with a variety of clustering parameters, and would conduct a task-oriented user test in order to evaluate the capacity of the structure to guide the user's exploration. The structure that performs best would be used online for mediation.

Similar applications can be offered to the users of an intranet and also to individual users, by using their hierarchic set of bookmarks as a source collection.

Internet search engines can also benefit from added functionality. Results of initial searches, based on general queries, can be clustered and used as a dynamic source collection. The user's actions such as selection of documents or clusters, can be used by the mediation system as a form of relevance feedback and better queries can be generated, explicitly or implicitly, for subsequent searches.

Mediation, as an extension to a retrieval system, should improve the support for typical user tasks. O'Day [OJ93], studying professional searchers, has identified 3 types of retrieval tasks:

1. monitoring a well-known topic or set of variables over time.
2. following an information-gathering plan specific to the task at hand.
3. exploring a topic in an undirected fashion.

The first one can be supported by the use of bookmarks in the source collection, identifying topics of interest. The user can go through the bookmarked clusters and documents and have the system generate and submit to the Web the appropriate queries in order to identify "what's new out there".

The other two tasks are supported by the hierarchical structure of the source collection, together with searching and visualization tools. The difference is in the amount of searching and browsing: a planned, analytical search is expected to rely more on searching based on representative keywords, while a non-directed search is expected to rely more on browsing and finding serendipitous information.

The "more like this" approach also suggests applying this paradigm to other media such as pictures, music and so on, for which expressing a content-based information need is problematic. For example, a large collection of pictures can be indexed automatically based on content (colour spectrum, texture, etc). A representative small sub-collection can be annotated, in order to support textual queries, and clustered or categorised. The user can then explore the sub-collection through textual queries and browsing. When relevant pictures are found, the system can generate a content description of the relevant items and extend the search to the full collection.

3.3 Objective of the thesis and contributions to research

3.3.1 Main objective

The main contribution of this thesis is to propose system-mediated access as an effective interaction model for Information Retrieval. Therefore, the main objective of the thesis is to prove the usability and the effectiveness of the new paradigm. This objective is captured by 2 hypotheses:

Hypothesis 1 (Usability) *System-based mediated access is a usable information retrieval paradigm.*

Hypothesis 2 (Effectiveness) *System-based mediated access through a clustered specialised collection can improve effectiveness over un-mediated searching on a target collection.*

The two hypotheses are relatively independent. Hypothesis 1 needs to be tested in an interactive setting, with subjects using a mediation system to conduct searches on a number of topics. What we are interested in is that the users accept the conceptual model, can follow an appropriate interactive strategy, and are satisfied with the results and with

the search session overall.

Hypothesis 2, on the other hand, requires a comparison of two search models: the baseline one, and the mediation one. In such an experiment we can assume that the interactive retrieval systems reifying the two models are usable and that the users are able to follow the prescribed strategy. What we are interested in is comparing the search effectiveness in the two cases either through real user experiments, or through simulations.

While we intend to test these hypotheses, it is important to observe that the effectiveness of mediation depends on the coverage and specialisation of the source collection, on the clustering algorithms and parameters used for structuring the source collection, on the interactive strategy employed and on the formula to generate topic models.

Although we expect Hypothesis 2 to be experimentally satisfied, the result is based on a combination of factors whose effects are confounded. Therefore, we intend to design and run a set of experiments to test the assumptions that mediation is based on and to observe the effects of various parameters and search strategies on mediation. The objectives of these experiments are described in the following two subsections.

3.3.2 Cluster hypothesis

Successful mediation through a specialised source collection assumes that the topical, semantic structure of the collection is clear, i.e. documents covering a certain topic are clearly separated from the other documents of the collection. This is usually the case if expert manual classification has been employed.

We have proposed the use of document clustering as a cheap and flexible alternative for structuring the source collection, so we need to verify the assumption that clustering is successful at separating documents into topics. While this assumption appears to be a consequence of the *cluster hypothesis*, we argue against its original formulation: "Closely associated documents tend to belong to the same clusters and to be relevant to the same request" [JR71]. Firstly, it mixes two issues: the capacity of a clustering method to group together similar documents (which depends on parameters and the actual clustering

algorithm employed) and the relationship between inter-document similarity and topical content (which is the cluster hypothesis proper). Secondly, it assumes a reciprocal, bi-directional relationship between similarity and relevance, which we have informally shown to be the case. We therefore propose to test the following hypothesis:

Hypothesis 3 (Aspectual cluster hypothesis) *Highly similar documents tend to be relevant to the same topic. However, documents relevant to the same topic may be quite dissimilar if they cover distinct aspects of the topic.*

and its consequence:

Consequence 1 (Cluster hypothesis consequence) *Clustering algorithms tend to group together documents that cover highly focused topics, or aspects of complex topic. Documents covering distinct aspects of complex topics tend to be spread over the cluster structure.*

The cluster hypothesis specifies the relationship between inter-document similarities and relevance based on topical content, and is independent of any clustering or classification method. Its consequence relies on the capacity of clustering (and the literature review has indicated that algorithms vary with regards to this capacity) to group together highly similar documents.

We expect clustering to support the exploration of a specialised collection by grouping together similar documents, so that relevant topics and subtopics can be identified. Moreover, we expect that 'good clusters' exist; these are, intuitively, clusters which contain a high number of relevant documents, and a comparatively low number of non-relevant documents. For each query, we expect most relevant documents to be contained in a few number of clusters, which contain few non-relevant documents. However, for complex information needs, with distinct aspects, the documents relevant to different aspects are expected to be grouped in separate pockets of relevance.

While we expect our cluster hypothesis and its consequence to hold, it would also be desirable to observe the contribution that the indexing strategy, the weighting scheme and the clustering method have on it.

3.3.3 Topic models

Unlike other research work in Document Clustering, we propose the use of multiple cluster representatives, each appropriate for a different purpose. To that end, rather than using heuristic methods found in literature, we adapt recent advances in statistical language models, which combine a strong theoretical model with flexibility to combine documents and clusters. We propose three hypotheses that express requirements for successful mediation:

Hypothesis 4 (Browsing labels) *Browsing labels convey content and can successfully guide navigation of the source collection.*

Hypothesis 5 (Searching labels) *Searching labels convey content and can successfully support search strategies in the source collection.*

Hypothesis 6 (Mediation labels) *Mediation labels can support effective search of the target collection.*

These hypotheses refer to the labels generated through the formulae described in section 3.2.5. We intend to verify Hypothesis 4 during the user experiments designed to assess the usability of the mediated approach. The other two experiments are intended to be tested through a combination of real user experiments and simulations.

Due to time constraints we are restricted to test simple formulae based on language models. Future experiments are envisaged for evaluating various smoothing techniques, for comparing our formulae with classic formulae from the literature [JR71, Voo85b], and also for comparing system-based mediation with various query expansion techniques.

3.3.4 Search strategies

We intend to compare the exploration strategies described in section 3.2.8. While we cannot confidently propose a hypothesis, we intuitively expect fuse-and-search to be a recall device and search-and-fuse to be a precision device.

We intend to compare search strategies in mediation simulations that look at:

1. absolute upperbound performance levels attainable by an 'ideal user' who can perfectly identify the relevant documents in the source collection.

2. upperbound performance levels relative to the cluster structure used for mediation attainable by an 'ideal user' who can perfectly identify the best clusters. These results, of course, depend on the algorithm and parameters used for clustering the source collection. It is unlikely that any cluster would contain all the documents relevant to a topic and no non-relevant documents. In other words the cluster hypothesis is not expected to hold perfectly. Therefore, the performance levels obtained through these types of simulations are expected to be more realistic than the absolute upperbounds. Real users are expected to be able to reach them if they follow appropriate exploration strategies and are able to recognize good clusters.

3. the potential improvement made by weighted queries, compared to unweighted ones, on the quality of the result.

4. the variation of the search effectiveness with the size of the query.

3.4 An evaluation framework

3.4.1 How to evaluate ?

In section 2.4 we have looked at current views with regards to evaluating interactive IR systems. We adapt Robertson and Hancock-Beaulieu's view of combining laboratory and operational tests [RHB92]. Their approach typically means finding the best components through laboratory tests, and then putting them together into an optimal system to be tested in an operational setting, with real users.

Our approach is slightly different, due to our peculiar context. We are proposing a generic solution, namely system-based mediation based on a structured specialised collection, that can be implemented in a variety of ways for a variety of specific situations or tasks. The software framework that we have built can be used to generate a variety of specific applications based on mediation, in which a variety of operation modes (opaque or transparent) and search strategies are supported. However, limitations in time and resources do not allow us to set up and run formal user experiments for several scenarios

and to compare the outcomes.

Therefore, we have chosen the following evaluation flow:

1. We start by evaluating the usability of the user interface and its ability to support mediated retrieval. During the iterative design-implementation-testing cycle we built several interfaces, based on different parameters (such as the clustering algorithms and the cluster representative generation formula) and supporting different scenarios. We present here conclusions from testing our first 'public' interface as well as the current one, named ClusterBook[8].

2. We conduct experiments on clustering the source collection in order to test our cluster hypothesis assumptions and to identify the clustering algorithms and parameters that are best at separating the topics of the domain. Apart from trying to show that clustering can support mediation by grouping together topical documents, we also intend to test our aspectual cluster hypothesis.

3. We run mediation simulations in order to confirm the usefulness of labels generated through language models, and to compare various search strategies in terms of retrieval effectiveness. The upperbounds of performance under various assumptions and using various strategies are expected to provide an indication of the theoretical performance achievable by the 'ideal user', and also guidelines as to which parameters and search strategies perform better. These guidelines can then be used in operational systems.

4. The final step is left for future work: operational systems specific for various tasks can be built based on the software framework and guidelines resulted from the work described in this thesis. They need to be tested in their operational context in order to assess whether they support users in conducting searches.

3.4.2 Experimental setting and test collections

A practical problem raised by testing the mediation concept is the availability of an appropriate source collection. In the operational scenario this would be a relatively small and homogeneous specialised collection which comprehensively covers the user's domain

[8]The name comes from the Library metaphor that supports it.

of interest (so that the user can find and select relevant documents in any of the domain's topics). We have already discussed that if no such collection was available, it could be built dynamically by searching the target collection with an initial query.

While the experimental setting should reflect the operational one, a perfect match is not always possible. In order for the source collection to support experiments to evaluate the multiple aspects of mediation, it should have the features of a test collection: a set of test topics, together with relevance judgements. Moreover, as we intend to explore the aspectual form of the cluster hypothesis, we need test topics for which distinct aspects have been identified and relevance judgements that distinguish between aspects.

During the development stages of the software a variety of specialised test collections were used in order to test the clustering algorithms, the cluster representative formulae and the user interface. As we are not reporting these informal experiments, no details of the document collections are needed[9]. We are only describing the test collections used in the experiments described in this thesis.

Reuters-21578[10], a collection of 21578 newswire articles from 1987, was the first collection selected for evaluating an IR system based on mediation, as well as some of the mediation assumptions. There are two reasons for selecting this collection:

1. Reuters is a test collection typically used for text categorization experiments. Its articles are manually annotated with names of categories or topics addressed by the articles. This supports experiments on the cluster hypothesis (testing how well documents that cover the same topic are grouped together) as well as simulations of the mediation process (estimating the quality of queries generated from documents known to cover a certain topic).

2. A subset of Reuters covering a short period of time can be viewed as a specialised collection covering the important world events that happened in that time interval. Such a subset of relatively small size (the Reuters collection is physically divided

[9]These document collections are available from a variety of sources such as http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/.

[10]http://www.research.att.com/~lewis/reuters21578.html

into segments of 1000 documents each, and we picked one segment) can be easily clustered and used as a source collection for mediating searches on the Web. Tasks for the users doing such experiments can be hand-picked by researchers from this source collection.

Unfortunately, the Reuters collection has no relevance judgements for narrowly focused topics, specific for user tasks, but only for rather generic categories such as "barley", "cotton", "fuel", "gold", "silver", "sorghum", "tea" and "zinc". Therefore, neither could the aspectual cluster hypothesis be tested on Reuters, nor could this collection be used as a test collection for retrieval effectiveness tests. It was useful, however, for testing the experimental software as well as the usability of our prototype.

For the formal mediation experiments we looked at the TREC (Text REtrieval Conference) experiments, organised by the National Institute of Standards and Technology (NIST), trying to find a more appropriate test collection. The Interactive track of TREC-8[11], was designed to investigate the exploration of complex information needs, with a multitude of aspects. This is the very situation in which a user would have problems formulating precise and comprehensive queries, and hence would find mediation helpful. Therefore, we decided to use the Interactive TREC-8 experimental design and the associated test collection, the Financial Times of London collection with 210,158 news articles from 1991-94.

Here are the six test topics associated with the TREC-8 experiment:

1. Number: 408
   Title: tropical storms
   Description: What tropical storms (hurricanes and typhoons) have caused property damage and/or loss of life?

2. Number: 414
   Title: Cuba, sugar, imports
   Description: What countries import Cuban sugar?

[11] http://www-nlpir.nist.gov/projects/t8i/t8i.html

Topics:	1 : 408	2 : 414	3 : 428	4 : 431	5 : 438	6 : 446
Aspects	24	12	26	40	56	16
Relevant documents	35	8	20	33	52	29

Table 3.1: Number of aspects and number of relevant documents for each of the 6 topics of the source collection.

3. Number: 428

   Title: declining birth rates

   Description: What countries other than the US and China have or have had a declining birth rate?

4. Number: 431

   Title: robotic technology

   Description: What are the latest developments in robotic technology and in its use?

5. Number: 438

   Title: tourism, increase

   Description: What countries have experienced an increase in tourism?

6. Number: 446

   Title: tourists, violence

   Description: In what countries have tourists been subject to acts of violence causing bodily harm or death?

plus another topic for practice:

- Number: 303

  Title: Hubble Telescope Achievements

  Description: Identify positive accomplishments of the Hubble telescope since it was launched in 1991.

Relevance judgements provided by NIST indicate not only which documents are relevant for each topic, but also considers aspects of each topic and indicates the relevance of documents to each aspect. In summary, **Table 3.1** shows the number of aspects identified for each topic as well as the number of documents judged relevant for each topic.

The FT collection used by TREC was chosen as target collection for our mediation experiments. As no specialised source collection was available, we simulated one based on

relevant documents from the FT collection. Based on the relevance judgements associated with the test collection, we split the relevant documents into two equal groups: one was included in the source collection in order to be used for mediation, and the other in the target collection, in order to be used for retrieval experiments. We attempted to cover each aspect of every topic with the relevant source documents, in order to better simulate a specialised document collection, which is expected to cover all topics of a domain. Typically a specialised collection contains more than just documents of interest to the user, so we made the source collection more realistic by 'polluting' it with copies of the 572 documents judged non-relevant. Due to the pooling system used by NIST to judge TREC documents, we know that these 572 'officially non-relevant' documents have been judged relevant and retrieved by some searchers, so they have some degree of similarity to the 'officially relevant' documents or some relationship to the test topics. The presence of these 'near-miss' documents, both in the source and in the target collection, is intended to make the experimental setting more realistic.

The source collection, although artificially built for our experiments, does have some of the characteristics of a specialised collection. It is relatively small (747 documents), so that it can be clustered and explored by a user through a combination of searching and browsing strategies, it has a relatively high concentration of relevant documents, and the majority of the documents, although not relevant, have some topical similarity with the relevant ones.

This experimental setting is sufficiently realistic to allow the testing of our assumptions and to support simulations that can offer a better understanding of the issues raised by using mediated access. Future mediation experiments with real users in operational settings are envisaged to use real specialised collections as 'sources' and the Web as 'target'.

Chapter 4

WebCluster - Design and Evaluation

4.1 Introduction

This chapter is devoted to the design and evaluation of WebCluster, our system that reifies the concept of mediated retrieval. We do not intend to cover in detail all the steps we took to build WebCluster, from user requirements through conceptual design, software design, implementation and testing. That level of detail is left to a technical manual.

We are interested in providing:

1. a *proof-of-concept* of the mediated retrieval concept. Without it, the mediation concept would be just an unproven hypothesis.

2. the reader with an idea of the look-and-feel, functionality and architecture of the software that we have built.

3. an insight into the functionality, flexibility and extensibility of our system to potential collaborators who may want to use it or further develop it.

In the following sections we first describe the architecture of the system, and its main components. The user interface deservedly receives the most attention as it is the component that supports the interaction with the user and thus the mediation process. As

mediation was discussed in detail in chapter 3, here we can concentrate on the actual system design. The second part of this chapter is dedicated to the evaluation of WebCluster.

4.2 General architecture

Mediated information retrieval relies on:

1. a set of source collections that cover various specialised domains, large enough to be representative, but small enough to afford exploration.

2. a structuring method that reveals the structure (subdomains, topics and concepts) of the collection and, implicitly, of the domain. The structuring of the source collection can be performed offline, for a static specialised collection, or on-the-fly, for a dynamic collection obtained through searching the target collection.

3. a highly interactive user interface with information visualization tools adequate for exploring the source collection through a combination of searching and browsing.

4. a model or formula that generates browsing and searching labels, and also produces good mediated queries based on selected clusters or documents.

5. a search engine for searching the target collection based on the mediated query.

For maximum flexibility we chose a *Client-Server* architecture, in order to separate the *presentation* aspect of the functionality from its *action* part. **Figure 4.1** presents the basic architecture of the system.

The Client is represented by the user interface or front-end through which the user interacts with the system, requesting services. It is lightweight and implemented in Java, which provides a high degree of platform independence.

Most of the processing is done by the Server: indexing of the documents, clustering of the collections, best-match and cluster-based searching of the source collections, generation of document and cluster representatives. In an operational system its main requirement is speed, so it was implemented in C++. It should run on a server with sufficient memory to take the inverted file of the source collection and to cache the document

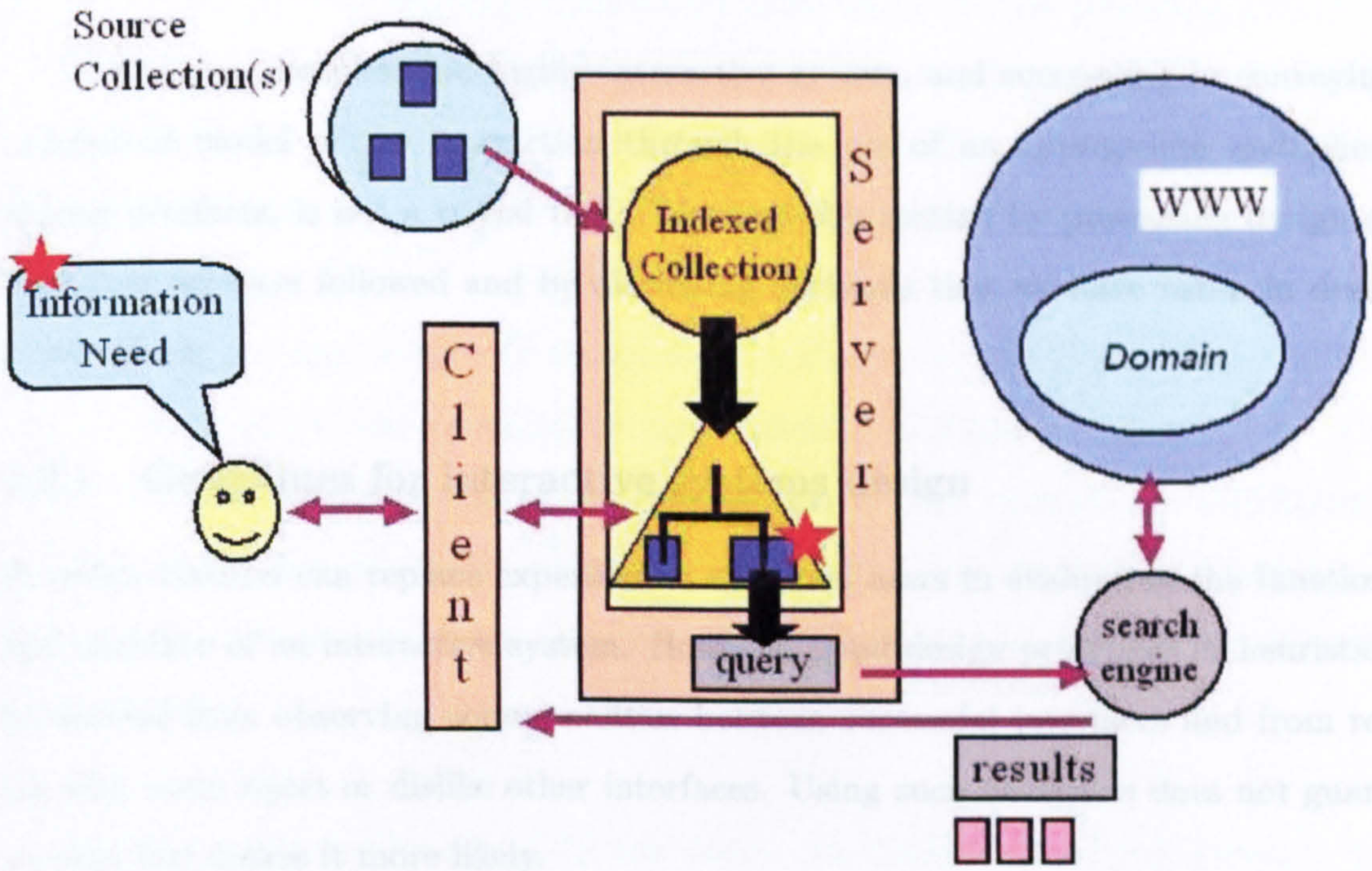


Figure 4.1: Basic WebCluster architecture.

and cluster representatives.

The Server's search functionality can also be used for searching the target collection, if that collection is available locally. If the target collection is the World Wide Web, then the mediated query is submitted to an external Web search engine. Although in principle any such engine will do, the WebCluster prototype uses Informia, a meta-search engine designed and built by Ubilab, our collaborators [BBMS98].

4.3 ClusterBook, the implemented user interface

While WebCluster is the name of the whole project, ClusterBook is the name of the interface that reifies the mediated access concept based on the library metaphor.

The design of the interface follows from the functionality required by mediation: the user explores the domain of interest represented by a structured source collection, selects documents or clusters representative for a certain information need, requests a mediated query from the system, edits the query if necessary, and submits it to a search tool on the

target collection.

Designing a complex and highly interactive system, and succeeding in conveying the conceptual model of the interaction through the use of an appropriate metaphor and display artefacts, is not a trivial task. We start this section by presenting design guidelines that we have followed and by discussing decisions that we have taken in designing ClusterBook.

4.3.1 Guidelines for interactive systems design

No other method can replace experiments with real users in evaluating the functionality and usability of an interactive system. However, good design principles or heuristics can be derived from observing commonalities between successful interfaces and from recording why users reject or dislike other interfaces. Using such heuristics does not guarantee success, but makes it more likely.

Sets of heuristics or 'golden rules' have been put together by various interaction gurus such as Norman [Nor88], Crawford [Cra92], Nielsen [Nie93] and Shneiderman [Shn98]. They prescribe various design choices that attempt to bridge the gap between the conceptual model of the system and the display artefacts in order to reduce the user's cognitive load, to prevent, spot and easily correct errors, and to make the interaction more efficient, effective and pleasant.

While attempting to respect most of them, we must be aware of some constraints:

- Our mediation system is envisaged to be a tool that offers extra functionality and better retrieval results to the user who is willing to read a short user manual or a tutorial. Therefore, we are not striving for the transparency of a walk-up-and-use system.

- We are building a research prototype, not a commercial product. Therefore, advanced interactive functionality such as shortcuts for the expert users, reversal of actions, help and documentation, or informative feedback have not been implemented, although they were considered during the design stage.

We employ *direct manipulation*, which gives the user a sense of control. We also try to prevent errors by only allowing actions that make sense and by providing default values for various parameters of the systems such as the source collection, the target collection, or the number of required hits. Probably the only possible user error, for which clear feedback is provided, is trying to do a search without typing in a query.

The artefacts and widgets of the system are clearly separated into panels of different colours, according to the functionality they provide (see **Figures 4.3** and **4.8**). A *colour model* is also used for feedback: selected documents and clusters are highlighted in a brighter colour, already visited documents are dimmed, and the synchronization between different views of the document collection is also supported by highlighting the common document(s).

The clearly marked functionality, the interaction through direct manipulation, the use of colour to mark the current and the already seen documents, and the cut-copy-and-paste functionality (for building a query) also cooperate in reducing the user's memory load and making the interaction a pleasant experience.

We also applied lessons learnt from successful IR interfaces[Hea99b]. Relationships between documents are conveyed by the hierarchic structure, while the relationship between the query and the search results is revealed by the meta-information associated with each hit: the user is shown the contribution of the query terms to each document being estimated as relevant. Moreover, the user's selection of search results highlights those documents in the context of the structured collection, so that the user can follow a foraging strategy.

If unable to formulate a query for searching the source collection, the user can employ the visualization tool provided and explore the structured collection. The mediated query, used for searching the target collection, is automatically generated by the system.

4.3.2 Design alternatives

Interaction determinism versus flexibility

As a problem solver, the searcher is setting goals, planning tasks, monitoring progress, examining solutions and optimising the solutions' quality. Hendry and Harper claim that an *opportunistic* style is better supported by an *informal information-seeking environment* [Hen96, HH97], where search techniques are represented with data-flow notation and where the searcher has control of the layout and is able to customise the system. The process of seeking information and solving a problem by satisfying an information need consists then in planning and managing the workspace and the information flow.

This approach may work well for expert searchers, especially after some training in the use of the system. For novice users though this is not the right design. Nielsen [Nie93, p.12] shows that novice users do not customise their interfaces even when such facilities are available. Even for expert users there may be problems:

- Users may not always make the most appropriate decisions; customisation is easy only if it builds on a coherent design with well-understood options to choose from.
- Different users can have very different interfaces, so the possibility to collaborate or get help is reduced.
- The customisation feature itself needs a user interface, which adds to the complexity of the system and to the users' learning load.

At least for the initial versions of ClusterBook we have decided to take a different approach and to design a deterministic interface that imposes mediation as the interaction model. Mediation is a new concept and it might be ignored by the users if it was not imposed through the interface design. We follow Kirsh's guidelines by attempting to simplify choice and perception in order to reduce the user's mental load and let the user concentrate on the task at hand and the strategies to be employed [Kir95]. We try to enforce the conceptual model of the system and help the user develop the appropriate mental model for the interaction.

An 'expert' version of the interface may be considered in the future, by integrating the mediated interaction in a flexible and non-deterministic *information seeking environment.*

Conveying the information space structure

In designing the user interface, the idea of presenting the user with the full structured source collection was challenged by alternative approaches. For example, if an ontology (in the sense of a structured system of categories) exists for a certain domain, it may be sufficient to present to the user just the category labels, rather than use a particular document collection. However, we decided that putting domain terminology in the context of documents may be beneficial for users that are new to the domain.

The metaphor for the interaction is also essential when considering how to offer navigation through the information space: we are modelling the physical library, rather than just an ontology catalogue. However, perhaps in a not so distant future the physical library may not be familiar enough to provide a good metaphor; perhaps the personalised digital library will be more familiar to the typical information seeker, so the assumptions for the designs of interfaces such as ClusterBook will have to be re-considered.

4.3.3 A look at the interface

Before going into the actual software design of the system we spent a lot of time sketching possible 'looks' of the interface and doing cognitive walkthrough based on the various scenarios that WebCluster was going to support. It is probably also appropriate for the reader to have a 'feel' for the system before we go into the design details. Therefore, we present two versions of the user interface and present some of its functionality through a *walkthrough* based on a search scenario. **Figure 4.2**, which shows the initial WebCluster user interface, helps illustrate the search scenario while **Figure 4.3**, which shows ClusterBook, the current user interface for WebCluster, highlights the main function panels.

Suppose a user is interested in finding on the WWW information on underwriting (in the banking context). The search would proceed as follows:

1. Use the **Source Collection Panel (1)** to select a source collection appropriate for the domain; for example, the Reuters collection of news articles might be the best,

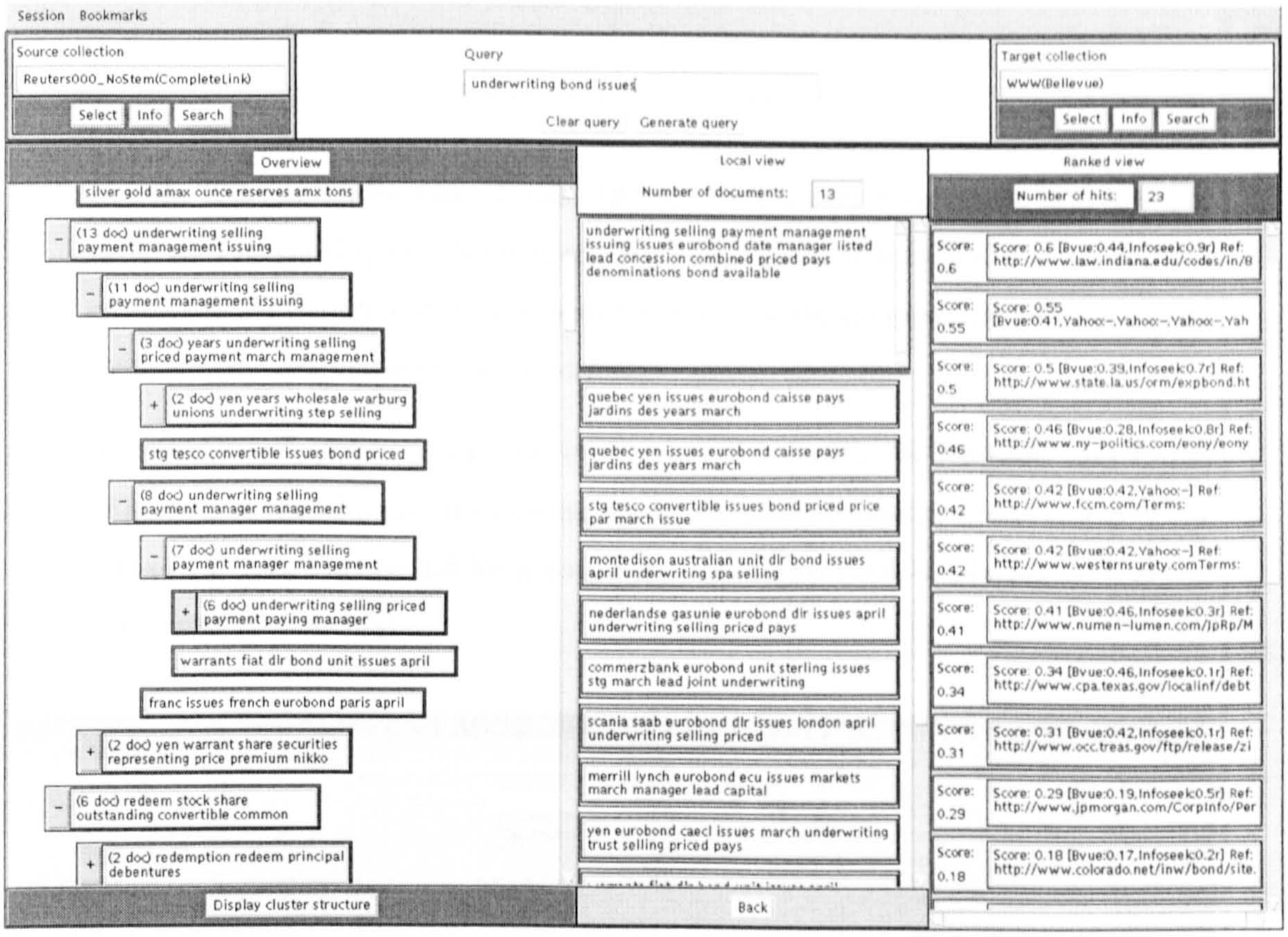


Figure 4.2: The initial WebCluster user interface, implemented with Java AWT.

among the available source collections, for a search for financial information.

2. Browse the clustered source collection, and find a cluster of relevant source documents corresponding to the information need. The clustered source collection is visualized in the **Overview Panel (4)**, and the hierarchy can be explored by the user by expanding/collapsing clusters. The **Local View Panel (5)** is used to display the subclusters of the selected cluster, or the content of the selected document. In our case, the cluster representative containing the keywords "underwriting selling payment management ..." attracted the user's attention and the exploration of the documents contained in the cluster confirmed the identification of a relevant cluster.

3. Ask the system to generate a query based on the keywords found in the cluster representative, derived from the source documents in the cluster. This query may be edited by the user, according to how precise or how comprehensive the search needs to be. In the **Query Panel (2)** the query has been edited and reduced to three terms, considered relevant and sufficient to express the user's information

need.

4. Use the **Target Collection Panel (3)** to select a target collection and query it using the generated query. Note, there may be a number of target WWW collections, corresponding to different WWW search engines; in the example the user has chosen Bellevue (the original name for Informia).

5. A ranked list of retrieved target documents are displayed in the **Ranked List Panel (6)**, and individual target documents can be selected for display in the **Local View Panel (5)** (see **Figure 4.8** for a screen-shot of ClusterBook displaying both source and target documents).

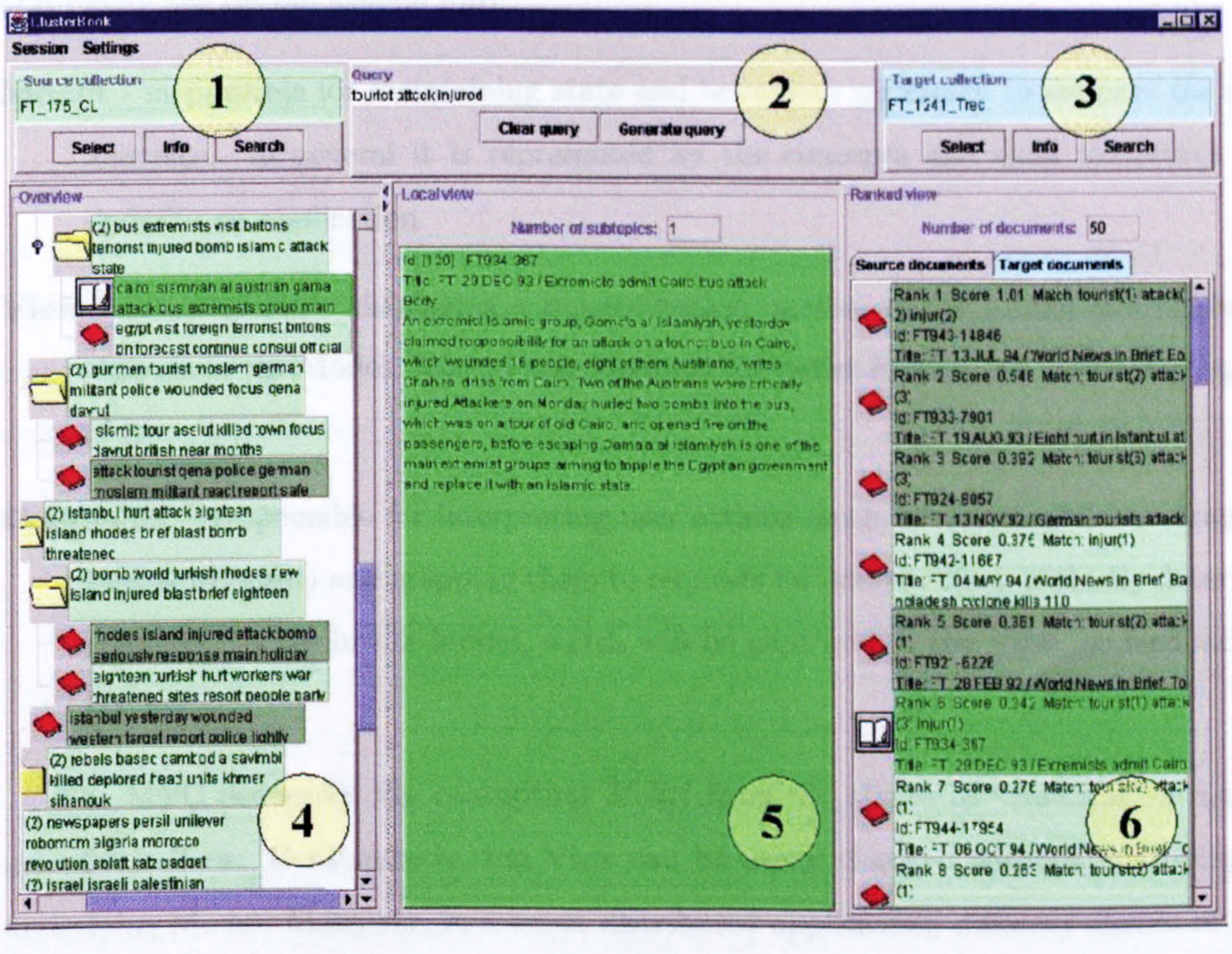


Figure 4.3: A version of ClusterBook.

We stress that this is just one of a multitude of mediation scenarios in which our system can be used. In the case where the user knows the domain and has no problems in formulating a query, it is also possible to search the source or target collection(s) directly,

by entering a query in the Query Panel and initiating a search of the chosen collection. The results are displayed in the Ranked List Panel (6).

4.3.4 The Model-View-Controller (MVC) framework

When designing WebCluster at the conceptual level, we decided to use the library as a metaphor for the overall interaction and the folder as a metaphor for the hierarchic structure visualization tool. However, we do not want to hardwire our software design: in the future we may find better interaction metaphors, or more suitable visualization tools.

A solution that offers flexibility and the possibility to easily change the implementation of the system is the Model-View-Controller design framework [BJ94, GHJV95]. It consists of dividing the overall system into:

Model - responsible for maintaining state and behaviour necessary to support the user-interface. In general it is represented by the concepts and data structures that underlie an application.

View - responsible for displaying a representation or view of the Model and conveying changes in the Model. It need not represent the entire Model, but only some aspects of it.

Controller - responsible for interpreting user actions (such as mouse clicks or drags, or keyboard input) and mapping them to requests for action. They typically determine changes of state in the Model, which will be captured in the View, as feedback for the user.

The MVC decouples the conceptual *Model* from the choice of visualization artefact used by a *View*. Consequently, the View can be easily changed without modifying the underlying Model. Moreover, in a large, distributed application, different classes of users can have separate Views of the common Model, full or partial, depending on their interest or task. A *Controller* is usually dependent on the View, as the user's actions are restricted to what is seen on the display.

Although the design and the implementation are distinct steps in building a system, the design needs to take into account the support that programming languages offer for

implementing the designed solution. A clever design is useless if it cannot be implemented effectively. This is a chapter about design, but we mention some implementation issues that have influenced the design.

The programming language selected for implementing ClusterBook was Java, due to its portability and the flexibility and power of its windowing toolkit, Swing [Gea99]. Swing's underlying architecture is built on the Model-View-Controller concept: the visual components which make up a view are 'connected' to a model and automatically change when the model's state changes. The Controller is integral to the visual components: the GUI events are interpreted, the appropriate requests are sent to the underlying model and the effects of the actions are reflected in the View (the display).

Of particular help in designing and implementing the visualization tool that supports navigation of the hierarchically structured source collection is Swing's JTree. It takes as model a tree-like data structure, which in our case is the hierarchy of documents and clusters and it deals with semantic GUI events such as branch expansion or contraction, selection of nodes and so on. It uses the Strategy design pattern [GHJV95] to delegate the rendering of the nodes to a TreeCellRenderer; the application programmer can replace or adjust the default renderer so that documents and cluster representatives are displayed appropriately.

Some of the above information may seem like implementation detail. In fact, it is quite the contrary: we argue that by selecting a programming language and a windowing toolkit that offer the appropriate support for implementing flexible solutions, we can concentrate on design issues rather than on implementation details. Moreover, a flexible design based on component interfaces, rather than implementation, offers the possibility of interchanging components and of easily adapting the user interface. For example, if Swing is extended with other visual components for displaying hierarchies, such as hyperbolic trees, tree maps or cone trees, it will be trivial to replace the JTree in a future version of the interface.

4.3.5 The Model

In this section we are looking at the conceptual model underlying our application, or in other words at the objects identified through the analysis of mediated search scenarios (discussed in chapter 3). The main classes of objects are described below:

Collection - It represents a collection of documents. It can be local or remote, linear or structured, accessible directly or through an access manager. The user can identify it by its name and can have access to its documents.

Document - It represents a document in the source or target collection. Although our current application only deals with textual documents, a flexible design allows the use of other media in the future.

Cluster (Category) - It supports the modelling of the hierarchic structure of a document collection, by recording the parent cluster and the children at each level in the hierarchy.

Vocabulary - Either independent or derived from the collection following indexing, it controls what terms can make up document and cluster representatives, and can be used for searching the source collection. In the current implementation we are only using words, but phrases or other representations such as n-grams could be used in the future.

Cluster representative - Conceptually, this is a vector of weighted terms. In fact we are using indices (from the Vocabulary) rather than actual terms. This improves the efficiency of the system, but also the flexibility: word representations can easily be replaced by other features. Documents can be viewed as singleton clusters, so cluster representatives are also used as document representatives.

Query - The query may be formulated by the user and typed in the appropriate input field, or can be generated by the system based on the user's selection of relevant documents.

Result collector - It is used for berrypicking relevant documents during the exploration of the source collection and the search of the target collection, in view of storing or printing.

Bookmark - It is used for decreasing the user's mental load. Although captured in the design, it has not been implemented.

System - Conceptually, this object represents the functionality of the system such as managing and searching the document collections, or computing cluster representatives. In practice, it does more than that: it realizes the *mediator* design pattern, loosely coupling the other objects of the application.

4.3.6 The Views (and the Controllers)

Different Views are appropriate for different scenarios and more than one View can realize a certain scenario based on a certain choice of the visual metaphor. However, due to the common goal, most of these have the same generic components. Below, we use the names of the components identified through the search scenario described in subsection 4.3.3:

Source Collection Panel - allows the user to select a source collection that covers a certain domain of interest.

Overview Panel - displays the hierarchic, topical structure of the source collection, obtained through clustering or categorization. The user can browse the structure by scrolling the panel, expanding and further exploring branches that look interesting, and collapsing branches that present no interest.

Local View Panel - shows details of the user's selection. It can displays the full content of a document or, alternatively, the full representative of a cluster, together with its immediate descendants.

Query Panel - allows the user to input a query, or the system to display the mediated query.

Target Collection Panel - allows the user to select a target collection, the search of which is the user's ultimate goal.

Ranked List Panel - displays a ranked list of hits, together with meta-information. The hits can be documents from the source collection or from the document collection, or can be clusters of the structured source collection, according to the search parameters set by the user.

If we chose to try and enforce the library metaphor more strongly, then specific library terminology could be used in the interface; for example, "Library Selection Panel" would probably be more appropriate than "Source Collection Panel". However, the names used here are quite generic and closely describe the conceptual model of the interaction.

The Controllers are tightly connected to the Views, as they deal with interpreting the user's GUI actions in the View; they are not a major design issue.

4.3.7 Dual access interface: multiple views in practice

ClusterBook is an interface that allows dual access to a document collection by combining

1. a hierarchic, structural view, based on the clustering of the collection, and a

2. a linear view, based on ranking the collection relative to a query.

The user has the choice of using best-match searching of the collection, which produces a list of documents ranked according to the estimated relevance to the user's query. However, the ranked view on its own fails to indicate the relationship between the retrieved documents, their commonalities, and the topical structure of the retrieved set.

The user can use cluster-based searching, in order to identify clusters of interest, or use the structure to

- browse the clusters of interest, looking for serendipitous relevant documents, learning the vocabulary of the collection, and getting more familiar with the collection.

- disambiguate terms and concepts, based on the context.

- assess the relevance of documents, based on the context.

The system supports the user in combining the two views of a document collection, hierarchical and linear, by highlighting in the overview panel documents selected by the user in the ranked view panel and vice-versa. Of course, if the user employs best-match searching in order to identify documents relevant for an information need, the distribution of these documents in the overview will indicate 'hot spots' of relevant documents and

will encourage the user to explore these spots by browsing.

ClusterBook can also be used as a research tool. An IR researcher can use it to visually verify the *cluster hypothesis*, by observing the distribution of relevant documents in the cluster structure or the distribution of the documents of a cluster in the ranked list of hits. Work with test collections (containing documents, queries and relevance judgements) is supported by colour-coding relevance judgements.

4.4 The server

4.4.1 The software architecture

Building an Information Retrieval system from scratch is not trivial. On the other hand, re-using and combining components from an IR toolkit or framework not only speeds up the process, but also offers flexibility, which is essential when iteratively designing, implementing, testing and modifying a new system. Moreover, the application builder can concentrate on the conceptual design and on conveying the conceptual model through the user interface, rather than on detailed software design and implementation.

In the WebCluster project we built not only a mediation system based on clustering, but a toolkit of components and a framework for building IR applications. In designing the IR framework we made heavy use of well known *design patterns* [GHJV95, CS95, VCK96, MRB98], which confers flexibility but also communicates the design in a language accessible to system designers. Therefore, building or extending applications based on the components provided in our framework is relatively easy.

Figure 4.4 illustrates the use of a combination of design patterns that offers flexibility in a variety of operations. The **Strategy** pattern allows a software client to delegate a semantic operation (e.g., clustering a collection of objects, computing the similarity between two objects, or computing the weight of a term in a document representation) to specialised objects that 'know' how to deal with that kind of operations. Flexibility is provided by a hierarchy of subclasses able to deal with a variety of cases, and the desired class instance can be selected at run-time. For example, in order to compute

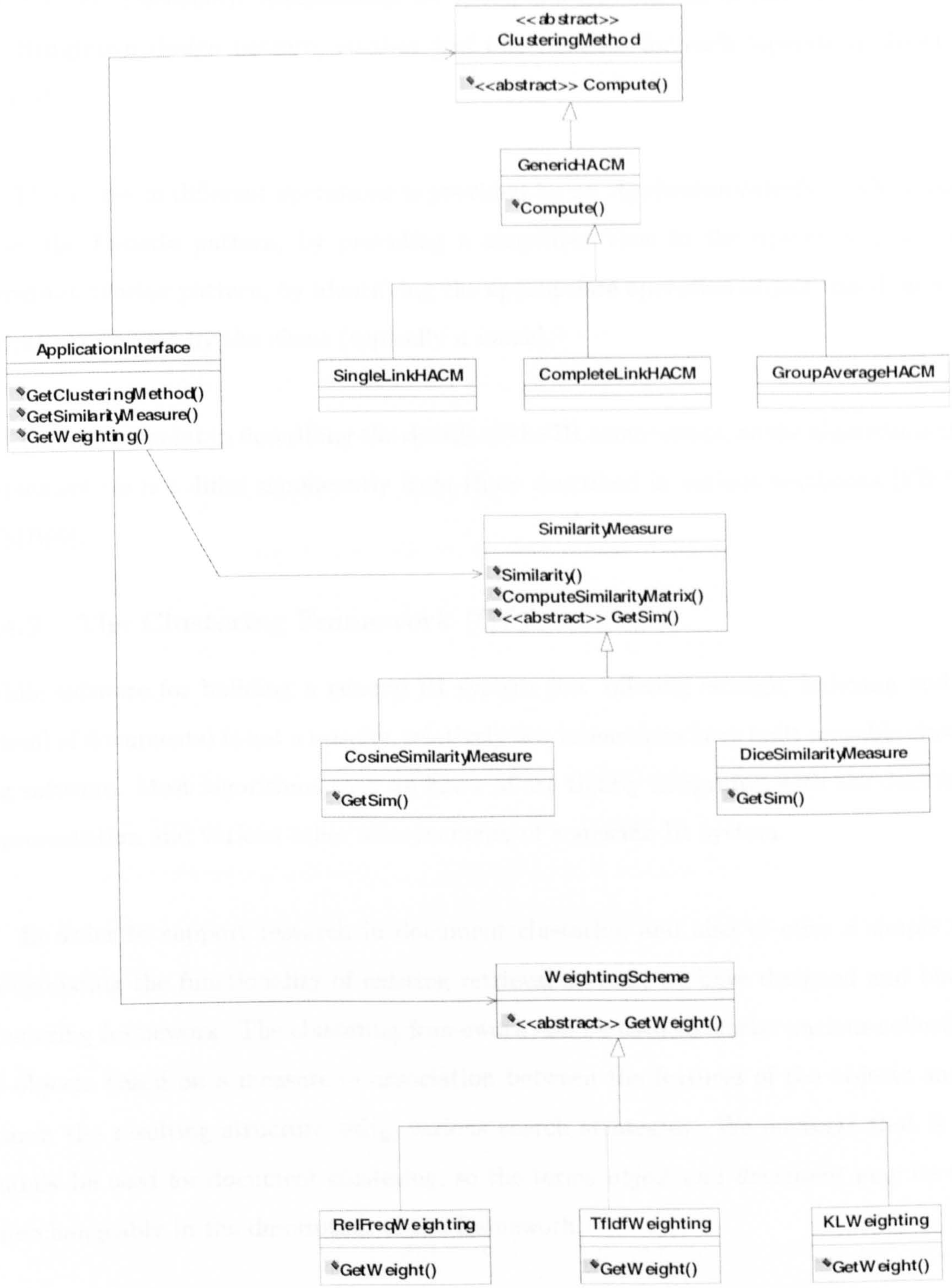


Figure 4.4: A flexible design of semantic operations.

the similarity between two objects, an application can use an instance of any subclass of a generic *SimilarityMeasure* class (we have implemented *CosineSimilarityMeasure* and *DiceSimilarityMeasure*). Additionally, for speed and efficient use of memory we also use the **Singleton** design pattern, so that just one instance for each 'operation object' is created.

The access to different operations is provided by an *ApplicationInterface*, which combines the **Facade** pattern, by providing a simplified view to the operations, and the **Product-trader** pattern, by identifying the appropriate operation object based on a description provided by the client (typically a name).

There is no point in describing the details of the IR components, as the algorithms they implement do not differ significantly from those described in various textbooks [FBY92, WMB99].

4.4.2 The Clustering Framework (CF)

While software for building a general IR system (i.e. offering storage, indexing and retrieval of documents) is not a novelty, relatively few researchers have built reusable clustering software. Most algorithms that we know of are tightly integrated with the document representation and various other idiosyncrasies of a specific IR system.

In order to support research in document clustering and also to offer a simple way of extending the functionality of existing retrieval system, we have designed and built a clustering framework. The clustering framework can be used to cluster various collections of objects based on a measure of association between the features of the objects and to search the resulting structure using various search strategies. We envisage that it will mainly be used for document clustering, so the terms *object* and *document* may be used interchangeably in the description of the framework.

Our design objectives were:

- Generality - it should work with different sources of documents, such as document collections or information retrieval systems.

- Flexibility - it should be able to use different clustering methods, similarity (or dissimilarity) measures, search strategies, methods of calculating the cluster representative, decision criteria and halt criteria.

- Extensibility - new modules should be added easily, so that the range of clustering parameters can be extended as new clustering algorithms, search strategies or similarity measures are implemented.

- Document independence - different document representations should be allowed for the collections to be clustered. The CF either uses already computed similarity values, when available, or converts the documents to an internal representation (based on the vector-space model), and then the calculation of the similarity between documents is done on the basis of the obtained representatives.

- Storage management independence - the CF should be able to use different storage media. This was achieved based on the **Serializer** design pattern [MRB98]. The initial implementation was based on ObjectStore, a proprietary OODBMS, which allowed fast implementation. In order to offer portability, we also added the possibility of using normal (Unix) files for storage.

A set of hierarchic and non-hierarchic clustering algorithms have been implemented, although only the former were used for mediation. An application programmer (or an end user, if the clustering framework is embedded in a flexible user interface) has at her disposal a range of options with regards to search strategies or the generation of cluster representatives.

4.5 The client-server interface

Conceptually, the Client and the Server are the two parts of WebCluster.

The Client is the user interface that realizes the interaction user-system and supports the user's information seeking tasks. However, the Client is just a lightweight user interface that contains visualization tools and controls which allow the system to interpret the user's actions and requests.

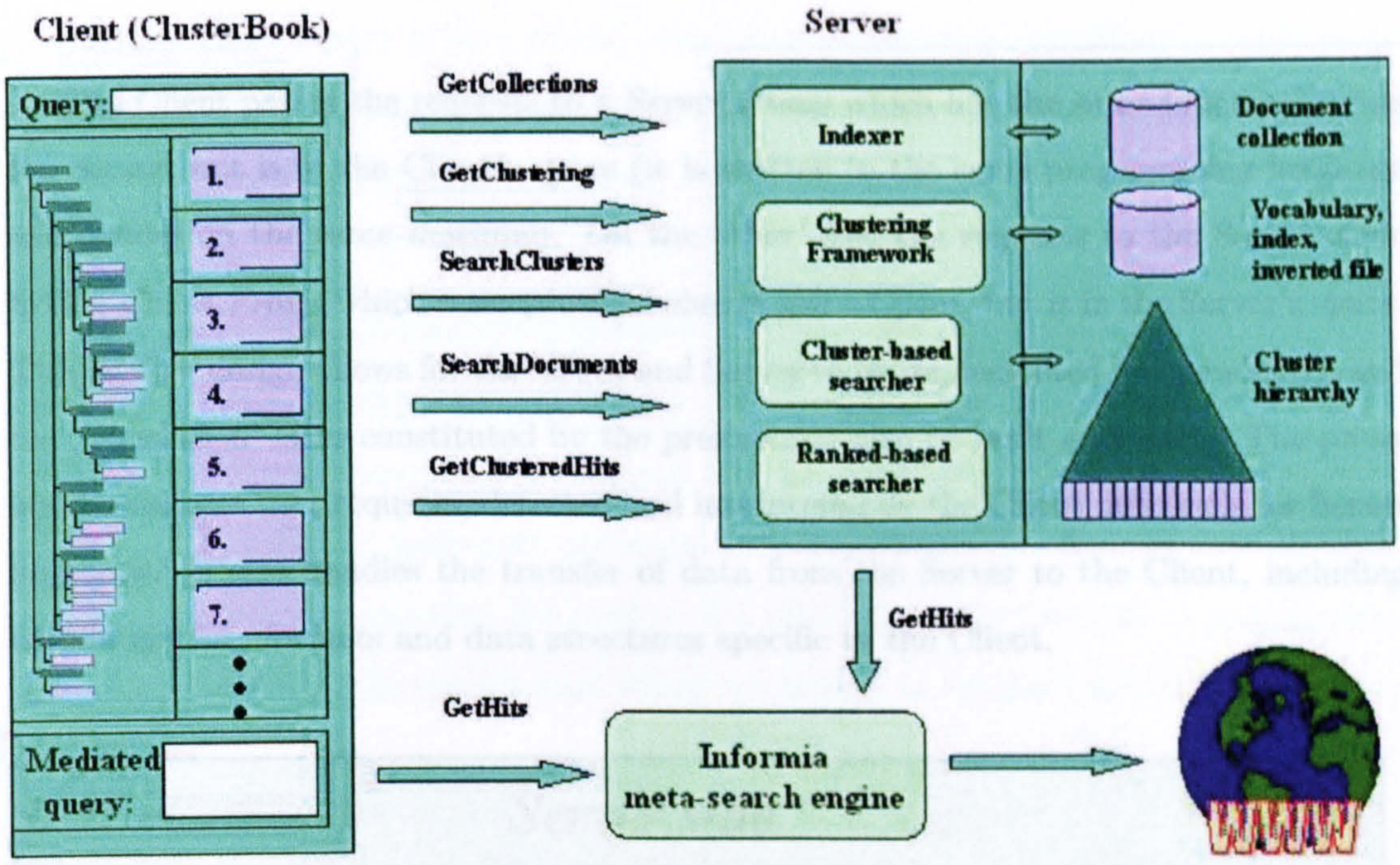


Figure 4.5: More detailed WebCluster architecture.

The 'hard work' is done by the Server, which is basically a toolkit of IR functions. It can also be used as an IR framework for building Information Retrieval systems. In WebCluster the Server's functionality is first used offline, by a system administrator, for indexing and clustering source collections. It is then used online, in the operational system, by integrating it into a process that waits for user requests. In response to requests coming from the Client, the Server responds by returning data: the hierarchic structure of a clustered collection, full documents or document representatives, cluster representatives, as well as results of best-match or cluster-based searching.

The Client and the Server are implemented in different programming languages (Java and C++) and they typically reside on different computers (a lightweight desktop and respectively a fast server). **Figure 4.5** shows the separation between the two entities.

In our design we intended to separate the conceptual relationship between Client and Server from the actual communication details. In order to achieve a high degree of flexibility and to keep the Client and the Server loosely-coupled, so that they can be inter-changed, we employed the **Proxy** design pattern [GHJV95], as shown in **Figure 4.6**.

The Client passes the requests to a *Server_Proxy* which has the same functionality as the Server, but is in the Client's space (it is written in the same programming language and resides on the same machine). On the other side, the requests to the Server come from a *Client_Proxy*, which conceptually behaves like a Client, but is in the Server's space. This flexible design allows for the Client and Server to be implemented independently, and the 'translation' layer constituted by the proxies can also be built separately. The proxy layer translates user requests, detected and interpreted by the Client, into calls for Server functions. It also handles the transfer of data from the Server to the Client, including translation into formats and data structures specific to the Client.

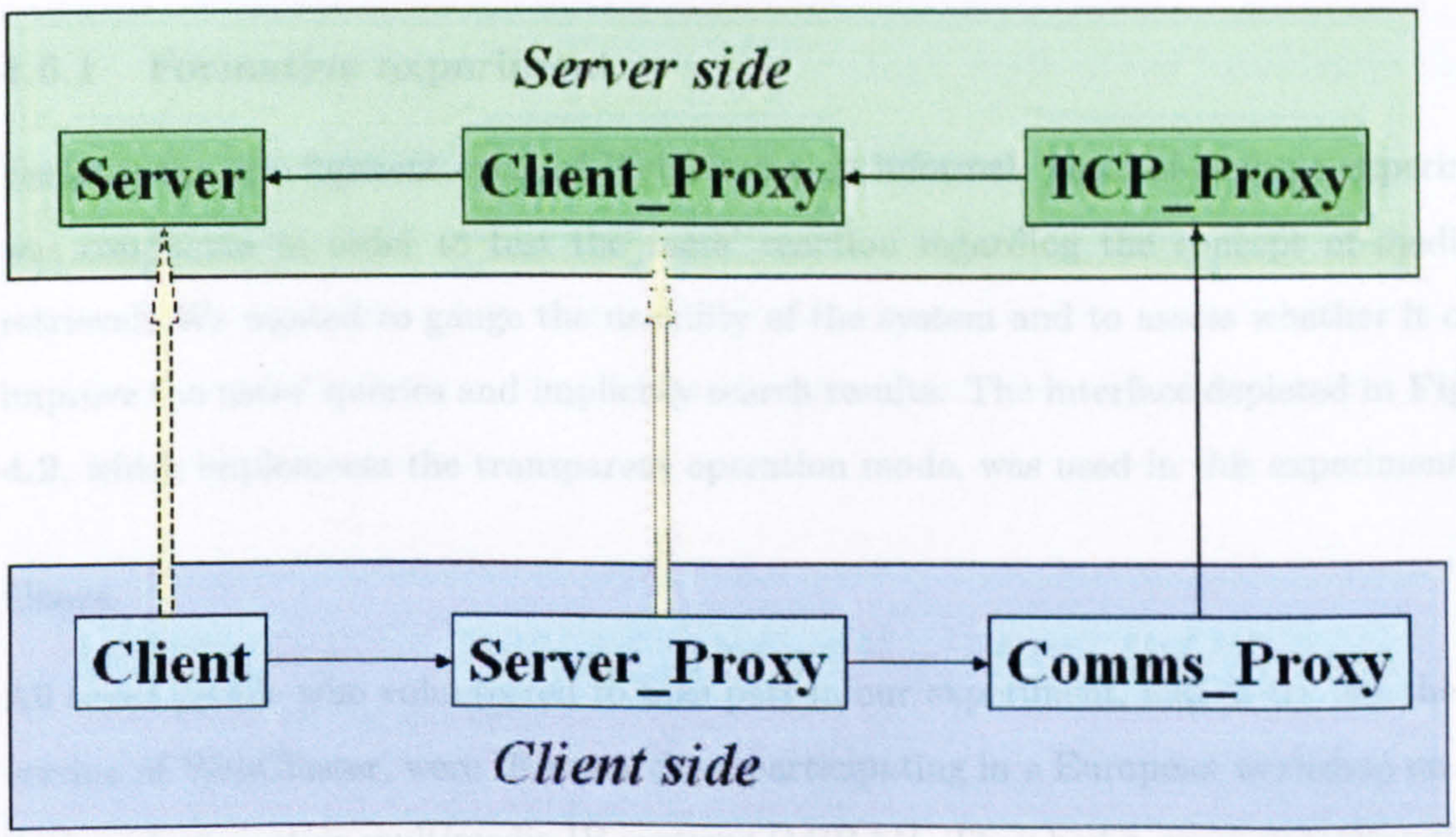


Figure 4.6: The client-server communication.

A further level of indirection is necessary in order to have independence of communication protocol. The Server_Proxy and the Client_Proxy delegate the actual communication to CommsProxies and have a choice of HTTP and TCP Internet protocols. The TCP implementation is faster and more reliable, while HTTP allows connections through a firewall.

4.6 Evaluation of WebCluster

Evaluating WebCluster is part of the evaluation framework described in section 3.4. We are particularly interested in the usability of the system and its capacity to enforce the mediation interaction model.

We first present a formative experiment that helped us identify and fix usability problems in the first version of WebCluster. Before another set of user experiments was designed and run, a user-less evaluation based on the analysis of the design space was conducted, in order to anticipate potential problems and to predict the success of the new interface. Finally we describe the results of a pilot study that we ran in preparation of an Interactive TREC - like experiment which we intend to conduct in the future.

4.6.1 Formative experiment

Early in the development cycle of WebCluster an informal, proof-of-concept experiment was conducted in order to test the users' reaction regarding the concept of mediated retrieval. We wanted to gauge the usability of the system and to assess whether it could improve the users' queries and implicitly search results. The interface depicted in **Figure 4.2**, which implements the transparent operation mode, was used in this experiment.

Users

All seven people who volunteered to take part in our experiment, and to try out the first version of WebCluster, were IR researchers participating in a European workshop on evaluation of interactive multimedia IR systems (MIRA)[1]. They had a good understanding of search strategies and indexing, and were experienced in using IR systems and in formulating queries. Their high level of expertise allowed them to not only follow the experimental procedure as end-users, but also to act as expert reviewers. Their comments and suggestions on the usefulness and usability of the system during the think-aloud retrieval sessions proved extremely valuable.

[1] http://www.dcs.gla.ac.uk/mira/

Procedure

1. The subjects were introduced to the idea of mediated retrieval and to the use of the system through a short tutorial.

2. The work task situation was described. The subject had to imagine that she was a journalism student on a work placement at a large national daily newspaper. Her job was to support the journalists in writing articles by finding relevant information on assigned topics. The sample tasks, manually selected by the experimenters from a collection of Reuters articles, were:

   - The journalist is writing an article on the coffee industry. She wants to know (if and) how quotas for growing coffee are set and controlled on a world-wide basis.
   - The journalist writing an article on strategic stocks of raw materials in the US wants to know details of the US oil reserves.
   - The journalist wants details on the history of the Brasilian debt crisis.
   - George Shultz has visited the Soviet Union for talks with Gorbatchev on a missile reduction programme. Details of the visit and of subsequent related visits are needed.

   The subject (journalism student) was expected to retrieve as much relevant information as possible, so that a journalist could find enough useful material to write an article on the subject.

3. The subject was asked to pick one topic (the recommendation was that the subject knew as little as possible about that topic) and to write down the query that she would be likely to submit to a search engine.

4. The subject was asked to select the Reuters collection from a pool of clustered source collections and to browse it[2] in order to find the best cluster that matched the description of the topic (she could explore the cluster representatives and also the documents that made up the cluster). The user could ask the system to generate a query based on the chosen cluster, and could edit the query before submitting it for a search on the target collection (the Web, indexed and searchable by Informia).

[2]Searching had not been implemented at the time of the experiment.

5. The subject used the initial, self-generated query, for a search on the target collection and compared the results to the ones obtained based on the mediated search.

There was no time limit imposed on search sessions. However, as they took place during a busy research workshop, the participants were trying to finish quickly, which probably reflected real information seeking situations derived from assigned tasks. The *think-aloud* protocol was used. The examiners took notes on the users' reasoning and actions, as well as on their comments and suggestions for improving the system. No logging of the actions and no formal post-task questionnaires were used.

Results

Most users felt comfortable with the idea of mediated access. They found it particularly useful when they were not familiar with the problem domain and therefore had problems in producing keywords for the query. This confirms our expectation that mediation can be valuable in assisting a user formulate an information need, especially during exploratory searches.

When the topic was more familiar, the users felt they could generate good queries without assistance and showed a certain resistance to spending time with the mediation stage. However, constrained to using the experimental procedure, they did use mediation and were pleasantly surprised to see the improvement in retrieved results due to this stage.

The resistance noticed corroborates with other research results showing most users to be result-driven rather than process-driven, trying to get just the minimum required of a task, with a minimum of effort [MDKL93]. This would suggest that the interface implementing the opaque operation mode, with only one retrieval stage, might be better received by the users than the one implementing the transparent mode, which involves a two-stage process. However, the user interface implementing the opaque scenario was not ready for testing, so a comparison was not possible.

Some users questioned the necessity of searching the target collection (the second step of the explicit scenario) when some documents retrieved from the source collection were relevant for the task topic. This may be due to the lack of an explicit stopping condition

for the search, respectively to a somehow imprecise task. It may also be due to a lack of rigour from subjects in assessing if the task has been achieved: after starting the search, they never re-read the task or compared it to the documents retrieved, neither did they consider if the information gathered would be enough for writing an article. No user made any attempt to find information on related topics. They wanted to stop as soon as a minimal set of results seemed to be sufficient for satisfying the task.

When the user edited the query proposed by the system, by deleting the terms that were clearly not relevant, a significant increase in precision and recall was noticed, compared to the search on the user's original query. However, when the query was not edited, the results were sometimes worse. This was a clear indication that the formula we used for generating the mediated query was of poor quality. Moreover, the users found that often the cluster representatives did not convey well the contents of the clusters and had to look at sample documents to judge the relevance of the cluster. The cluster representatives and the derived mediated query were made up from non-trivial words that appeared in a high percentage of the documents in the cluster selected and no weighting scheme was employed. A better formula was clearly needed.

The users considered that there were too many clusters at the top level and it was difficult to distinguish between them. The problem was exacerbated by the lack of any search functionality. Hopefully best-match and cluster-based searching would alleviate this problem in the future.

Many keywords recommended by the system were observed to improve retrieval, and this provides some evidence of the benefits of mediated search for assisting query formulation. However, users often rejected proposed terms that were unfamiliar to them, even though some of these terms were known to the experimenters to be relevant for the task. This observation corroborates with research in *interactive query expansion* that has revealed inexperienced users' failure to recognize 'good' terms proposed by the system [MR97]. Therefore, if the transparent scenario is to be successful, tools will have to be provided to assist the user in making good decisions e.g. by showing terms in context or by ranking the list of proposed terms [RPH+95].

The users pointed out some lack of functionality in the system, such as a *colour model* for labelling different kinds of objects (e.g. documents already seen), an integrated *browser* for viewing the WWW documents, a *search* facility for the source collection, *bookmarks*, a *results collector*, a *history* function, and so on.

Discussion

Although it highlighted some usability problems (which were solved in the next version of the interface), the experiment was successful in that it indicated the potential of mediation. The users seemed to like the system and the mediated query did tend to improve the search effectiveness when compared to the users' initial query.

The experiment did not prove however, that a system based on mediation is superior to a baseline system (which simply offers query-based searching and visualization of the retrieved documents) at supporting a user perform a typical information-seeking task. In the time used for exploring the source collection in the mediated search, a user of the baseline system could have reformulated the query in order to improve the retrieved set. Although our comparison was somewhat realistic (as most Web searchers do not follow the initial search with successive queries [JSBS98]), a fairer comparison would be between a system supporting mediation and a baseline system which allows the user to view the retrieved documents and to reformulate the original query.

4.6.2 Design space analysis

Following the feedback from the formative experiment described in the previous section, based on the first version of WebCluster, we improved the design and implemented a new user interface. The improvements in the system addressed:

- the functionality of the system, by adding best-match and cluster-based retrieval and by synchronizing the hierarchic and the linear (ranked) view of the source collection, so that selections in one view are reflected in the other.

- the usability of the system, by using a consistent colour model to separate source from target documents, to distinguish function panels, to reinforce the synchroniza-

tion between the views, to highlight selections and to dim clusters and documents already visited.

- the efficiency of the system, by introducing multi-threading and by optimising the transfer of data between the client and the server, and the caching of data.

We used cognitive walkthroughts [PRSB94] and gave demonstrations of ClusterBook in which we followed, step by step, the recommendations of the tutorial and of the user manual. No more usability problems were highlighted.

We have also adapted the *Cognitive Dimensions Framework* proposed by Green and Petre and successfully used by Hendry and Harper [HH97], in order to analyse the 'design space' of ClusterBook:

closeness of mapping - Our interface follows quite closely the library metaphor. Following a short tutorial, most users in our formative experiments have shown confidence in using the interface. Further testing would reveal whether changing the terminology of the interface (e.g., using a "Choose library" button label rather than "Choose collection") could improve the users' perception of the interaction model.

diffuseness/terseness - ClusterBook attempts to maximize the information density of the display in order to convey the structure of the information space, the estimated relevance of documents to the information need and document details. For large source collections, even the complete use of the screen may be insufficient to simultaneously display the structure of the collection, and document details. Possible improvements could be *zoom* functionality for the Overview Panel, or moving the Local View Panel, used mainly for viewing the document, into an independent frame.

role expressiveness - The overview quite successfully conveys inter-document relationships and the topical structure of the source collection by clustering together similar documents. The query-document relationship is apparent from the Ranked View Panel, and some term-document relationship is also shown by indicating, for each retrieved document, the query terms that contributed to it being retrieved. Highlighting query terms in documents is a possible improvement, as is a colour-coded system similar to TileBars [Hea95].

secondary notation - The layout of the workspace and the use of colour have no influence on the execution of services, but make a huge difference in comprehensibility. Feedback from users indicate that they strongly appreciated the introduction of the colour model, even if the choice of colours was not to everybody's taste.

side-by-side-ability - The structured overview of the source collection and the ranked list of hits can be viewed simultaneously, which offers great help for exploration. The system does not allow to view simultaneously two documents in order to compare them, but we envisage to try such a feature in a future version of the system, in which documents can be visualized separately from the navigation window.

viscosity - This dimension addresses the support for the user to easily repeat, with different parameters, a task already conducted. The current version of ClusterBook has no such provisions: it is highly deterministic and viscous, in order to better support more casual users. A more flexible version of the system, in which the user could describe the steps of a search strategy and run them for various topics or various source collections, may be built in the future, targeted at expert users.

While the interface has high viscosity for the end-user, it is highly flexible for the application programmer. This is due to the approach we have taken in order to build the system, by starting with the design and implementation of a toolkit and framework. It is possible to change with ease the layout of the interface, the position or colour of various panels, but also parameters such as the similarity measure or the clustering method used for structuring the collection, or the weighting scheme used for searching.

Although aware of possible future improvements, we are satisfied with the usability of ClusterBook, as indicated by principles of good design and cognitive walkthroughts, and with the general functionality of WebCluster. The rest of this chapter will now look at preparations for a future TREC-like user experiment.

4.6.3 The Interactive TREC-8 pilot experiment

In section 3.4 we justified our plan to use the model of the Interactive TREC-8 experiments for testing our mediated retrieval system. For reasons that will become clear, only a pilot user experiment was run, rather than the full experiment. However, for completeness and

Subject	Block 1	Block 2
1	System 1: 6-1-2	System 2: 3-4-5
2	System 2: 1-2-3	System 1: 4-5-6
3	System 2: 2-3-4	System 1: 5-6-1
4	System 2: 3-4-5	System 1: 6-1-2
5	System 1: 4-5-6	System 2: 1-2-3
6	System 1: 5-6-1	System 2: 2-3-4
7	System 2: 6-1-2	System 1: 3-4-5
8	System 1: 1-2-3	System 2: 4-5-6
9	System 1: 2-3-4	System 2: 5-6-1
10	System 1: 3-4-5	System 2: 6-1-2
11	System 2: 4-5-6	System 1: 1-2-3
12	System 2: 5-6-1	System 1: 2-3-4

Table 4.1: Block design that controls for the effect of topic and topic order.

better clarity, we present the full experimental methodology, together with the outcome of our endeavour.

Methodology

The experiment was designed to compare the experimental system (WebCluster) against a baseline one. They are based on the same indexing, weighting scheme and search algorithms, only differing in that the experimental system offers mediation. The expectation is that mediation should help the user formulate better queries and obtain better search results.

As there are 6 test topics, a Latin square design imposes the use of minimum 12 people, each doing 6 searches, 3 on the experimental system, 3 on a baseline system. **Table 4.1** offers an example of such a design, which controls for search topics and the order in which the users do the searches. The 12 subjects of the experiment are randomly allocated a number (from 1 to 12) and consequently the topics and the order in which these topics have to be addressed. For example, subject 1 must first conduct searches for topics 6, 1 and 2 (in this order) on the baseline system, and then for topics 3, 4 and 5 on the experimental system. It is apparent that users 4, 7 and 10 have the same groupings of topics, but the order or the allocation to the systems is different.

The task of the searcher is to save documents which, taken together, contain as many

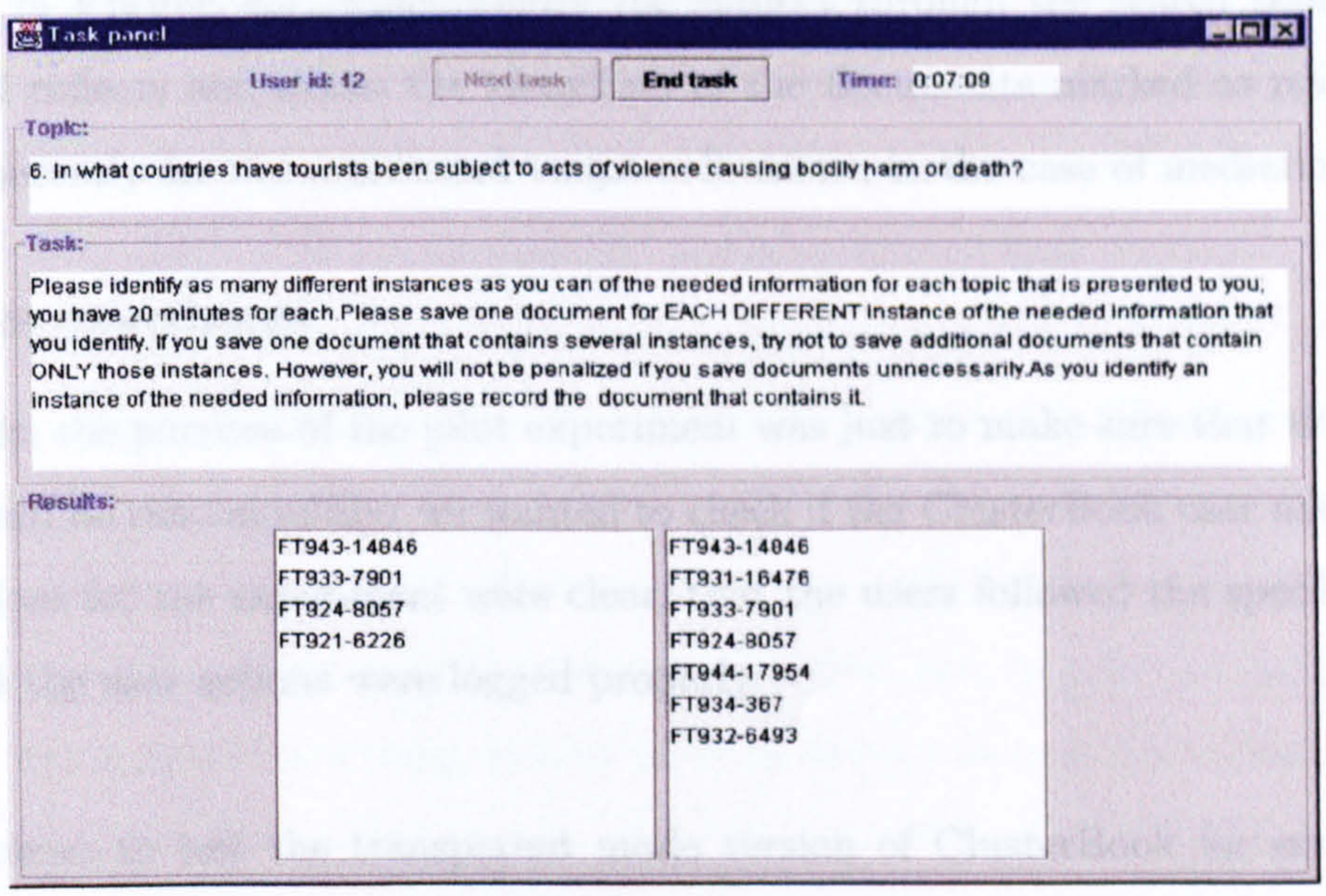


Figure 4.7: The Task Panel.

different instances as possible of the type of information the topic expresses a need for - within a 20 minute time limit. Searchers are encouraged to avoid saving documents which contribute no instances beyond those in documents already saved, but there is no scoring penalty for saving such documents.

More methodological details for this experiment are publicly available from NIST[3], so there is no point in reproducing them here. They include, among others, instructions to be given to the subjects and questionnaires to be administered before and after the experiment, before and after using each of the two systems, and before and after each individual search.

NIST's recommended measures of effectiveness for this experiment are *instance recall* (the fraction of total instances for each topic that are covered by submitted documents) and *instance precision* (the fraction of the submitted documents which contain one or more instances). In order to make the results comparable to other experiments, and especially to mediation simulations we have also added *average uninterpolated precision* and *R-precision.*

[3]http://www-nlpir.nist.gov/projects/t8i

In order to support the user during the experiment we provided a "Task Panel", depicted in **Figure 4.7**, which guides the subject through the search tasks, keeps the time and collects and stores the identifiers of the documents marked as relevant by the user (separately for the source and target collections, in the case of mediation).

The pilot experiment

Originally, the purpose of the pilot experiment was just to make sure that the full experiment could be run smoothly: we wanted to check if the ClusterBook user manual and the instructions for the experiment were clear, that the users followed the specified scenario, and that the user actions were logged properly.

We chose to test the transparent mode version of ClusterBook for several reasons. Firstly, we assumed that by offering the user more insight into the mediation process we could better convey the conceptual model, and by giving the user more control over the process the user satisfaction would be higher. Secondly, we wanted to check the users' reaction to the mediated queries and to see whether any editing of these queries would occur. Therefore, this version of the user interface was implemented and tested extensively, in order to be operational.

Another experimental design decision was to have the user interface enforce a "search and fuse" strategy by restricting the user to select only one document or cluster before asking the system to generate a mediated query. The reason is that intuitively we expected this strategy to be a precision device, which would make the user concentrate on various aspects of every topic, and generate precise queries for each aspect.

Four graduate students in Computer Science performed searches on the mediation system and reported that the mediation hindered, rather than supported their exploration of the assigned topics. Moreover, the mediated queries generated very poor retrieval effectiveness on the target collection. The reasons for this result became clear when we analysed the interaction in some detail.

It appears that we have a situation similar to that described in section 3.2.4. It hap-

pens that at least some of the test topics represent minor features in the structure of the source collection, so the relevant documents are scattered throughout the structure and are part of clusters representing some different major feature. For example, documents about tourists attacked in Pakistan are part of a larger cluster about tourism in Pakistan. It is therefore difficult, if not impossible, to find a cluster that accurately represents an aspect of the user's topic. The consequence is that the mediation system cannot produce queries to support the search task. For example, the searcher is diverted towards aspects of tourism in Pakistan rather than helped to find more documents about attacks against tourists. The "search and fuse" strategy, which we had thought would allow the system to build accurate models of various topics of interest and to generate mediated queries that produced precise searching, proved to be disastrous in such a situation.

The failure of the "search and fuse" strategy, at last for the particular test collection and test topics that we used, does not imply the failure of mediation. When we encouraged the searchers to change the mediation strategy and to freely explore the source collection and to generate a 'mediated query' based on terms learnt from relevant documents in the source, they were able to improve their search performance.

Figure 4.8 exemplifies this process. Based on an initial, rather vague query ("tourist attack"), combined with browsing, a user found relevant documents covering different reports of attacks on tourists. She extracted topical terms from these documents and manually expanded her query (e.g., "tourist attack injured militant extremist police wounded"). By submitting this query to the target collection, she obtained a ranked list of documents in which more top-ranked documents were relevant than if she had submitted the initial query.

Moreover, we explored the structured collection and found several well-focused topics for which the prescribed mediation strategy seemed to work well. However, they were not among the 'official' topics and had no relevance judgements assigned, so we could not run formal experiments with them.

An analysis of other research groups' results in the Interactive TREC-8 experiment

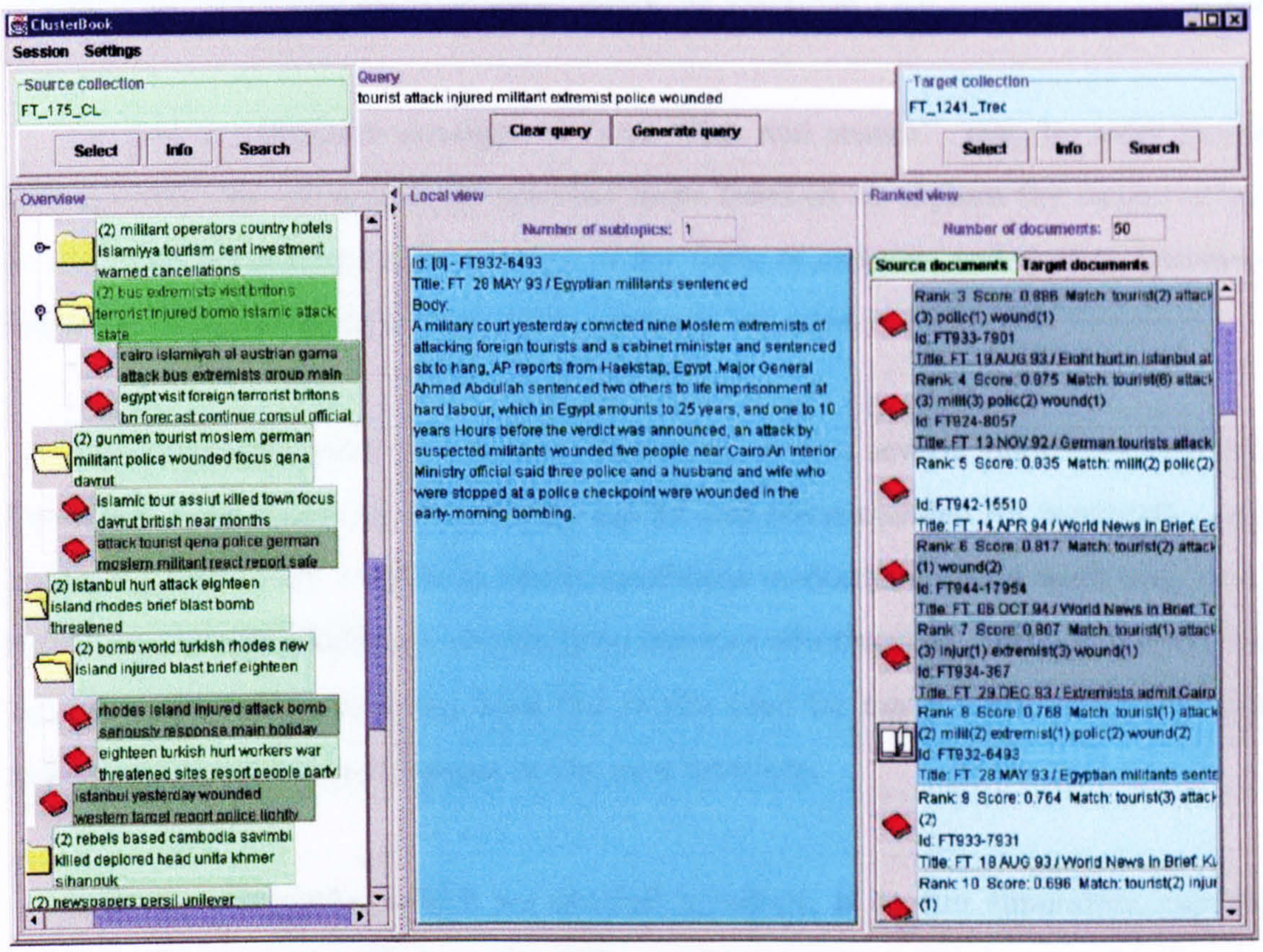


Figure 4.8: ClusterBook at work.

shows that most of them were disappointing: in most of these experiments an experimental system employed some technique designed to improve search performance, compared to a baseline system, but failed to do so [HO99]. A tentative explanation would be that the test topics are unfortunately of rather poor quality. They are vague and general and the relevance judgements were produced accordingly. Therefore, it is quite difficult to improve a basic query derived from the topic description, as it would lose the generality of the topic.

Hard decisions

Our initial intention had been to run an 'Interactive TREC'-like experiment as part of this project and to report the results in this thesis. However, the analysis of the pilot experiment indicated that a full experiment based on the transparent mode of operation and on a "search and fuse" mediation strategy would most likely be unsuccessful. Moreover, it would most likely not contribute to our understanding of the interactive information

seeking process based on mediation. In our view, this is due to the fact that the test topics available are too vague and therefore cannot be identified very successfully by clustering.

A different mediation strategy, such as "fuse and search", may be more successful. Alternatively, we could give the searcher more freedom to explore the source collection, to learn the terminology and concepts of her topic of interest, and thus to become more competent at generating a query that expresses her need.

One alternative that we had was, therefore, to run several pilot user experiments, for different combinations of operation modes and recommended (or imposed) mediation strategies, and to try to infer in which conditions mediations would work and, even better, in which cases mediation was likely to improve effectiveness. Such an approach would be extremely time consuming, both due to the need for many users, and to the need to implement the necessary changes in the user interface.

A better alternative, which we decided to adopt, is to run simulation experiments instead. Their purpose is to verify if the mediation assumptions are valid, to compare various strategies and operation modes, and to assess the influence of various parameters on the mediation outcome. This kind of simulations would allow us to establish *upper-bounds of performance*, attainable by the 'perfect user' in 'perfect conditions', but also to estimate the expected performance for the 'average user' in realistic circumstances. Consequently, user experiments can be set up and run in the future based on the best combination of parameters and strategies, as established by these simulations.

Another advantage of this approach is that, once in place, the software for simulations can be used repeatedly, on various test collections. Firstly, that offers the opportunity to compare the effect of collection characteristics on the effectiveness of mediation. Secondly, it can produce guidelines as to which combinations of parameters and strategies are more effective, in view of future user experiments, and of producing operational systems. The rest of this thesis deals with building an evaluation software framework and conducting simulation experiments.

Chapter 5

Clustering Experiments

5.1 Justification

This chapter investigates the potential of document clustering as a tool for structuring source collections for the purpose of mediated information retrieval. We are mainly interested in verifying the *cluster hypothesis* assumption of mediation, that clustering groups together similar documents and has the potential to identify topics and sub-topics. We also want to investigate a consequence of it: the potential of clustering to reduce the information space that needs to be explored by a user in order to find documents covering a certain topic.

We do not assume that all clustering methods will work on all document collections. Firstly, an attempt should be made at understanding the effect of various clustering parameters on the resulting clustered structure. Secondly, we envisage that for an operational system a system administrator would run a set of tests on the candidate source collections in order to validate the ones that are good for mediation. In this chapter we are essentially developing and running such a set of validation tests.

Some tests on the original cluster hypothesis are described in the literature review. We are adapting them in order to reflect our aspectual version of the cluster hypothesis.

Taking into account our use of the clustered structure for mediation, some limitations can be imposed on the range of the experimental variables, so that we concentrate on

what is relevant and potentially useful. Let us have a look at what may be accepted as a good structure in the context of exploration and mediation, and consequently establish which clustering algorithms we should investigate:

1. The structure needs to be intuitive. Therefore a hierarchic clustering method is probably more suitable than a partitioning method as it better reflects the reality (highly specific topics are included in more general topics) and common classification systems. Partitioning algorithms have proven very useful in grouping search results into meaningful topics, but they are less likely to be successful at supporting the exploration of full collections. Therefore, we will not consider them in our experiments.

2. The structure should be easy (and ideally pleasant) to navigate. Very deep and very wide structures increase the user's cognitive load and make orientation difficult. Also, seeing a high number of *aberrant* documents, which do not belong to any cluster, is very irritating for a user who browses the structure. The single-link clustering algorithm, which produces deep structures and a high proportion of aberrant documents proved very unpopular in informal user experiments, so it was eliminated from subsequent experiments. More popular were the complete-link (CL) and the group-average (GA) algorithms. The former generates small, tightly bound clusters of highly similar documents in the lower levels of the hierarchy, with good potential to identify coherent topics and sub-topics in the collection. It also generates a balanced structure, with few aberrant documents. The downside is a rather wide structure, with many branches at the top of the hierarchy; moreover, the structure tends to be rather random and meaningless high in the hierarchy, at lower level of similarity, as the algorithm looks for differences rather than commonality among clusters. One way to overcome this problem is by providing a search tool to identify starting points for browsing. The group-average clustering algorithm tends to achieve a balance between depth and width and consequently a general purpose structure. It is also expected to have a better recovery characteristic, i.e. to produce a more 'natural structure', with similar documents close to each other in the hierarchy.

3. Documents that cover the same topic should be grouped together, so that users

can easily recognise and identify topics and, if new to the domain of the collection, learn the topical structure of the domain with ease. The complete-link and group-average clustering algorithms also compete in satisfying this criterion. The former is expected to do better at the bottom of the hierarchy, by identifying small clusters of highly topical documents, while the latter is expected to give a better overall structure of the domain. Maybe a better solution is to use a combination of clustering methods: complete-link at the bottom levels, in order to identify topics, and group-average or single-link higher in the hierarchy, in order to identify connections between topics. Such mixed approaches to clustering may be explored in the future.

Despite the wealth of literature on clustering, the effect of clustering is still not well understood. Most experiments looked at the global effect of various clustering algorithms on the effectiveness of cluster-based retrieval, which is not necessarily useful for us. Even experiments that looked at the grouping of relevant documents in the clustered structure are less useful, as *aspects* of relevance have been ignored, presumably due to the lack of appropriate test collections. In conclusion, we still do not understand exactly the effect of clustering on grouping similar documents and identifying topics.

Clustering works by attempting to group together documents that have some degree of similarity, usually with regards to their term frequency distributions. Intuitively, documents that are highly similar are expected to cover the same topic and consequently clustering is expected to reveal the topics of a collection by grouping together similar documents. Therefore, our set of tests has two stages, in which we investigate:

1. the collection classifiability - estimate the potential of the document collection to be structured. We will investigate the **aspectual cluster hypothesis** (section 3.3.2), i.e. will be looking for a correlation between documents covering the same topic and being highly similar.

2. the collection clusterability - evaluate the **cluster hypothesis consequence** (section 3.3.2) i.e. estimate the quality of the structure built by concrete clustering algorithms in terms of grouping together topical documents.

Because clustering uses as input inter-document similarities, usually based on lexical content and on statistical analysis of term distribution in the documents, it is quite obvi-

ous that the classifiability of a collection is a prerequisite for its clusterability. If there was no correlation between the semantic, topical commonality between documents and their lexical content similarity, clustering could not be expected to group together semantically similar documents and to identify topics.

When planning our clustering experiments, we had to remember that they were part of the bigger picture we were building on mediation. Therefore it was the source collection used for mediation that we had to do the clustering experiments on. As explained in section 3.4.2, this collection was artificially built to simulate a specialised collection: it contains 175 documents judged relevant for at least one aspect of the 6 topics of the Interactive TREC-8 test (picked so that they offer coverage of all the topics' aspects), and 572 documents judged non-relevant; this gives a total of 747 documents.

It is worth stressing that this procedure of building the source collection is in no way biased in favour of ensuring good experimental results. On the contrary, we are trying to set up a realistic situation and even to make our life difficult by including the 572 documents that are non-relevant, but more or less similar to the topic descriptions. If clustering is successful at identifying topics in such a difficult situation, we would expect it to do much better when applied to more balanced source collections, where the similarity between relevant and non-relevant documents is lower.

5.2 The separation test

5.2.1 Experimental design

The cluster hypothesis separation test was proposed in its original form by van Rijsbergen and Karen Sparck Jones [RSJ73]. The experimental hypothesis is that there is a correlation between documents being relevant to the same queries (or topics) and being highly similar. The test calculates, for each topic for which relevance judgements are available, the similarities between each pair of relevant-relevant (RR) documents and each pair of relevant-nonrelevant (RNR) documents and checks that the average RR similarity is larger than the average RNR similarity. Additionally, the frequency distribution of the RR and RNR similarities are plotted and compared. The interpretation is that the less overlap

Topics:	1 : 408	2 : 414	3 : 428	4 : 431	5 : 438	6 : 446
Aspects	24	12	26	40	56	16
Relevant documents	35	8	20	33	52	29

Table 5.1: Number of aspects and number of relevant documents for each of the 6 topics of the source collection.

between the two histograms, the better separation between relevant and nonrelevant documents and the higher the collection classifiability.

We adapted this test for the *aspectual cluster hypothesis* that we have proposed, which conjectures that, while highly similar documents are expected to cover the same topic, documents that cover the same topic may be rather dissimilar if they focus on different aspects of the topic. To test our hypothesis we used the source collection of 747 Financial Times documents. The relevance judgements that come with the test collection were used to establish which documents were relevant to which of the 6 topics and, moreover, to which aspects of each topic.

Table 5.1 indicates, for each of the 6 topics of the test source collection, the number of distinct aspects and the number of relevant document, according to relevance judgements produced by human experts. The relevant documents are not uniformly spread over the aspects. **Figure 5.1** shows, for each topic, the frequency distribution of documents per aspect values. For example, for topic 1, one aspect is covered by no documents, 16 aspects are covered by one document, 4 aspects are covered by 2 documents, one aspect by 4 documents, one aspect by 6 documents and another aspect by 12 documents. Most aspects are only covered by just one document and a few aspects are not covered at all. However, a number of aspects are represented by more than one document, so there is just sufficient data to test the aspectual cluster hypothesis. The number of aspects covered by each document also varies. **Figure 5.2** shows, for each topic, the frequency distribution of the number of aspects per document. For example, for topic 1, 26 documents cover just one aspect, 8 documents cover 2 aspects, and one document cover 4 aspects. Most documents only cover one aspect (this is valid for all the documents relevant to topic 6), but a significant number of them cover several aspects and, in the case of topic 2, even all the aspects.

Our initial approach was to simply extend the original van Rijsbergen - Sparck Jones

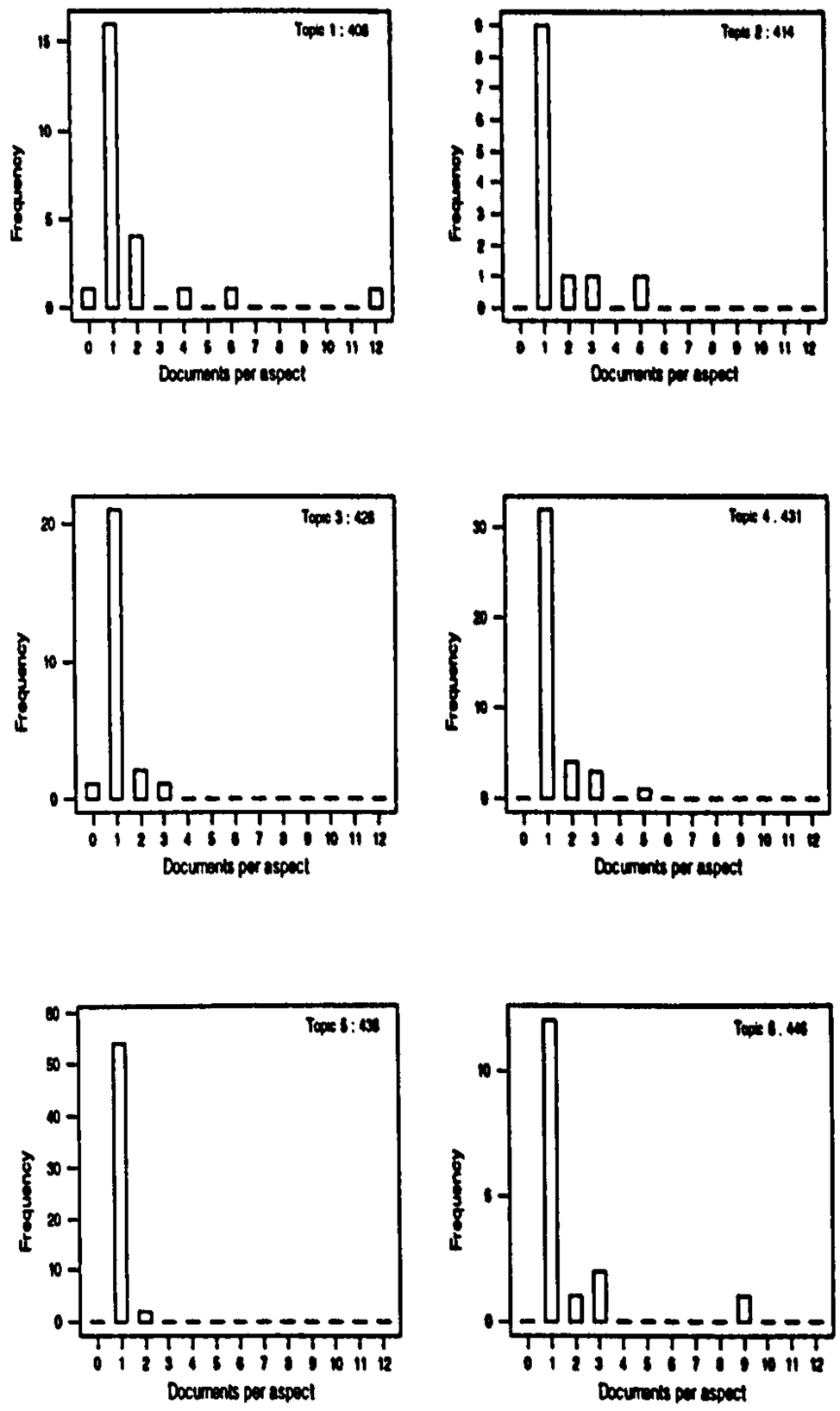


Figure 5.1: The frequency distribution of the number of relevant documents per aspect.

test by also including aspectual - aspectual and aspectual - non-aspectual similarities. However, while designing the algorithm to run the experiment an interesting issue came to light which can make the results counter-intuitive: imagine two documents, d_1, only relevant to topic T_i, and d_2, relevant to both T_i and T_j. The value of similarity between d_1 and d_2 contributes to both RR average (based on T_i) and RNR average (based on T_j). There is no problem if, like in the original experiments, the topics are very different, so that documents are unlikely to be relevant to more than one of them, and/or if the number of non-relevant documents is large compared to the number of relevant documents. This problem is more serious at the more detailed level of comparing similarity between pairs of documents that are relevant to the same aspect of a topic with similarity between

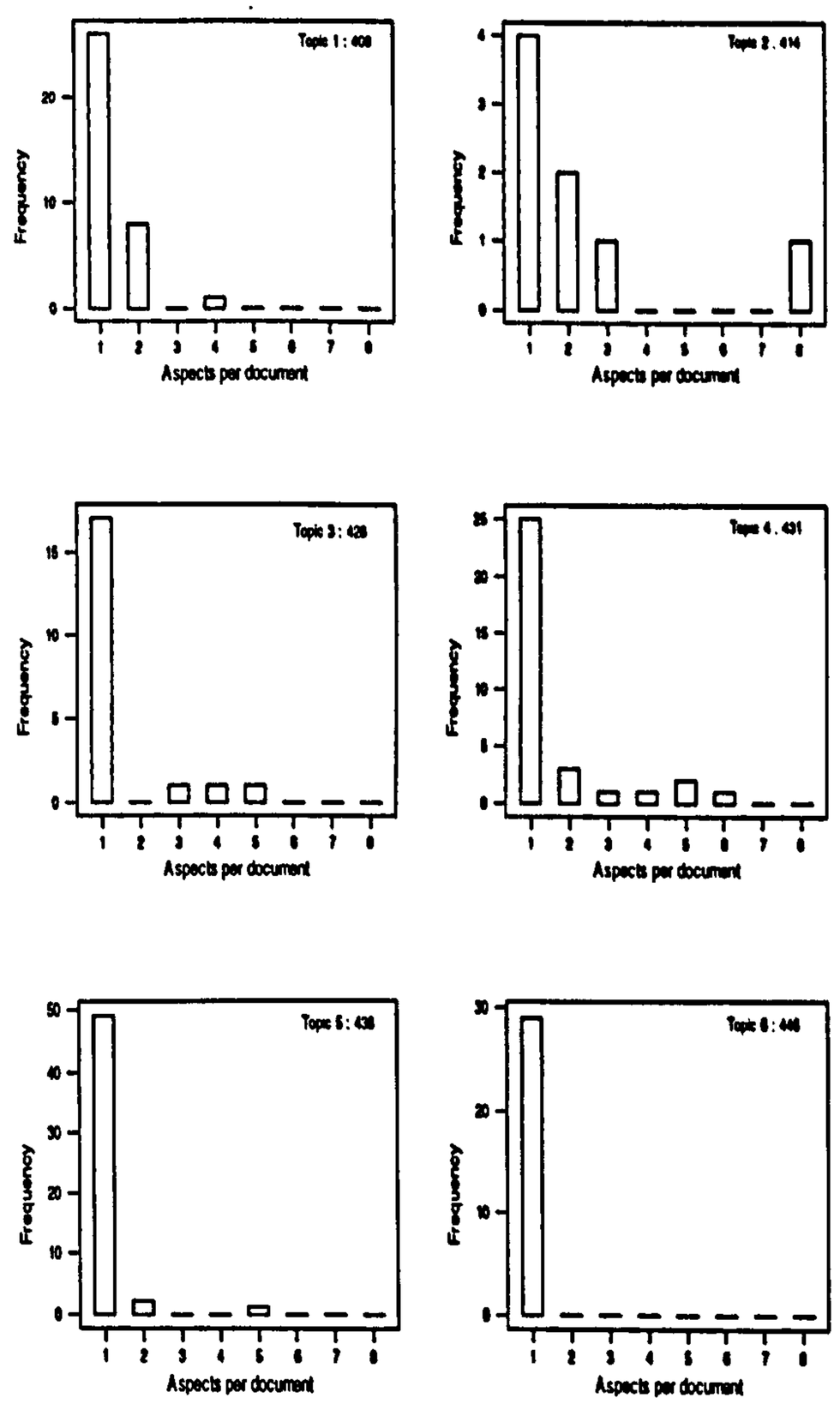


Figure 5.2: The frequency distribution of the number of aspects per document.

pairs of documents that are relevant to different aspects of the same topic. Due to the overlap between aspects of each topic and to most documents being relevant to more than one topic, applying the original algorithm gives the counter-intuitive result that pairs of topical documents that cover different aspects are, in average, more similar than pairs of topical documents overall.

In order to avoid this problem, we applied a simplified algorithm. We calculated the similarities between each pair of documents ("All similarities"), between each pair of documents relevant to the same topic ("Topical similarities") and between each pair of documents relevant to the same aspect of a topic ("Aspectual similarities"). The expecta-

tion is that aspectual similarities should be, on average, significantly higher than topical similarities, which should be significantly higher than all similarities. The computations were repeated with two similarity measures, *Cosine* and *Dice* (described in section 2.3.5), and three different weighting schemes for document terms: *relative frequency* (term frequency over the number of tokens in the document), *tf-idf* (in the Inquery form) and the *Kullback-Liebler (KL)* measure of divergence (as described in section 2.2).

5.2.2 Distribution of similarity values

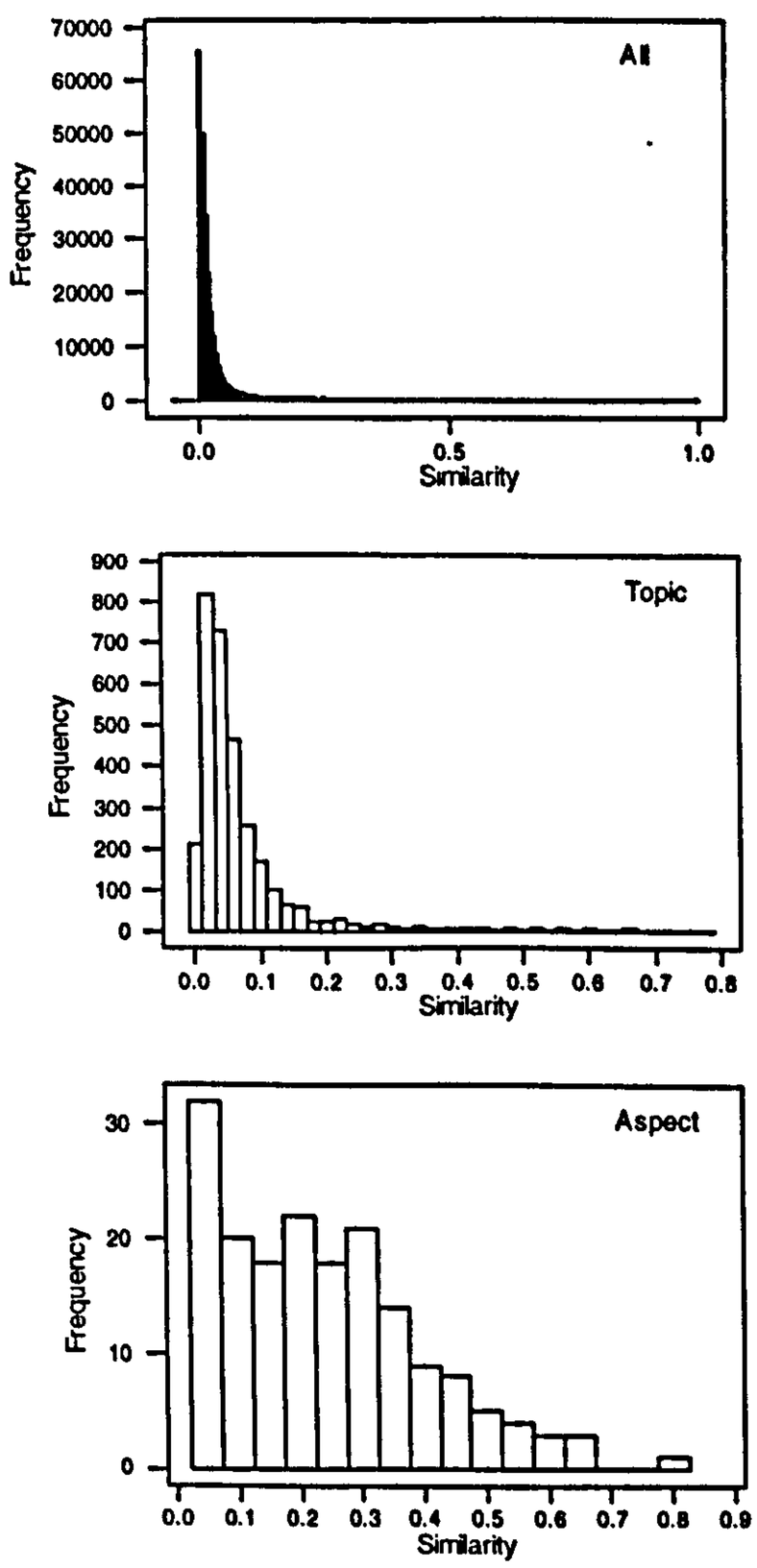


Figure 5.3: Comparison in histogram form of the distributions of all similarities, topical similarities and aspectual similarities.

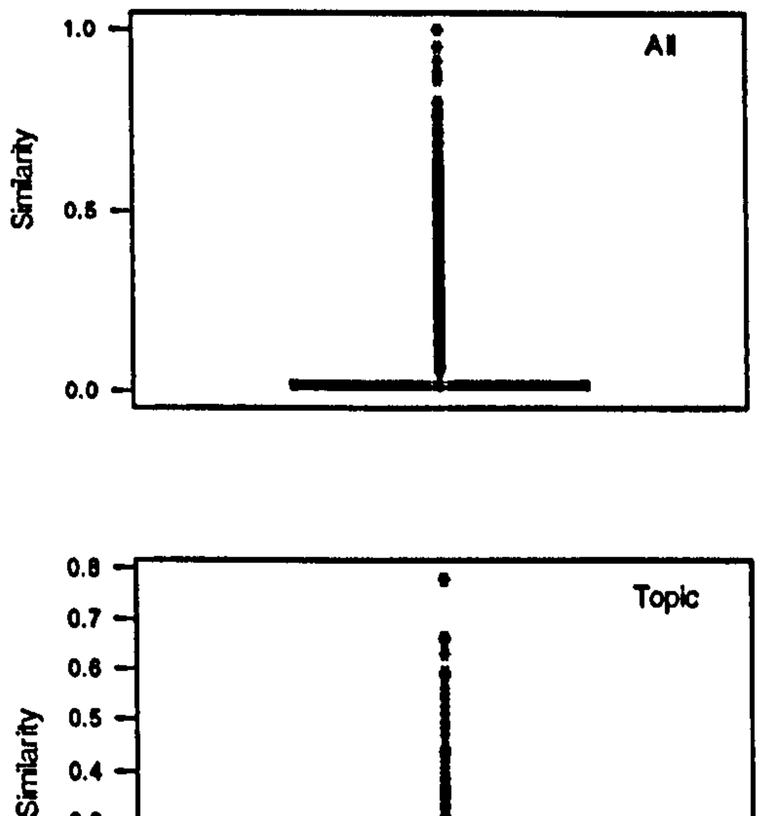


Figure 5.4: Comparison in boxplot form of the distributions of all similarities, topical similarities and aspectual similarities.

The histograms in **Figure 5.3** show the distribution of the 278631 "all similarities", of the 3073 "topical similarities" and of the 178 "aspectual similarities" for the Cosine-KL combination. They are all positively skewed, showing that relatively few pairs of documents are highly similar. The boxplots in **Figure 5.4** give less details on the distribution of similarity values, but better indicate the median and the interquartile range of the distribution, making the comparison between the three sets of values easier. It is clear, from examining these figures, that the experimental results confirm our predictions: documents that cover the same aspect of a topic tend to be more similar to each other than documents that cover the same topic, which tend to be more similar that randomly

selected documents in the collection. The results bellow, obtained with the Minitab statistical package, show that the separation between all similarities, topical similarities and aspectual similarities is highly significant[1]:

Descriptive Statistics: All, Topic, Aspect

Variable	N	Mean	Median	StDev	SE Mean
All	278631	0.02145	0.01184	0.03723	0.00007
Topic	3073	0.06474	0.04309	0.07474	0.00135
Aspect	178	0.2391	0.2199	0.1586	0.0119

Variable	Minimum	Maximum	Q1	Q3
All	0.00000	1.00000	0.00534	0.02415
Topic	0.00000	0.77701	0.02380	0.07430
Aspect	0.0254	0.7770	0.1078	0.3333

One-way ANOVA: All, Topic, Aspect

Source	DF	SS	MS	F	P
Factor	2	14.0898	7.0449	4869.38	0.000 < 0.01 (!!)

Individual 95% CIs For Mean
Based on Pooled StDev

Level	N	Mean	StDev	------+-------+-------+-------
All	3E+05	0.02145	0.03723	*
Topic	3073	0.06474	0.07474	*
Aspect	178	0.23915	0.15859	(*)
				------+-------+-------+-------
				0.070 0.140 0.210

While the *tendency* stated by the cluster hypothesis is shown to hold statistically, it is worth noting the outliers shown in **Figure 5.4**: there are a small number of pairs of documents that are not relevant to the same topic and still are highly similar; in fact some

[1]For the reader less familiar with statistical results, the most important values, which indicate the statistical significance of the results, are marked with (!) for statistical significance or (!!) for high statistical significance. The CI represents the confidence interval.

are more similar than all the pairs of documents covering the same topic or aspect of a topic. A plausible explanation is that we only took into account relevance for the 6 test topics, while the collection obviously covers other topics. It may well be the case that these highly similar documents are relevant to the same topic, but to a topic that had not been highlighted. Additionally, these outliers may also be due to human error, to judges failing to spot the relevancy, for the test topics, of some documents.

Another important results is that a significant number of pairs of topical documents have the similarity close to zero (or under a rather low threshold). This cannot be attributed to human judges wrongly assigning relevance to non-relevant documents. The most plausible explanation is that offered by our *aspectual cluster hypothesis*: documents that share a common topic may be very dissimilar if they cover different aspects of the topic.

The results presented above are just for one combination of similarity measure (Cosine) and weighting scheme (KL). Similar outcomes were obtained for all combinations of similarity measure and weighting schemes, so we can draw some definite conclusions for the collection on which the test was applied:

1. As suggested by van Rijsbergen's original cluster hypothesis, documents that are relevant to the same topic do tend to be more similar to each other than to other documents.

2. Furthermore, our prediction that the similarities between documents relevant to the same aspects of a topic are significantly higher that the topical similarities is confirmed.

5.2.3 Effect of independent variables

In the previous section we were interested in the theoretical outcome of testing the traditional and the aspectual form of the cluster hypothesis. Therefore, we were interested in showing that the results were consistent over the range of independent variables considered.

		RelFreq	TfIdf	KL
Cosine	All	0.074392	0.024604	0.021449
	Topic	0.197109	0.072660	0.064739
	Aspect	0.324279	0.241734	0.239146
Dice	All	0.067566	0.022215	0.017733
	Topic	0.186124	0.067183	0.056067
	Aspect	0.298656	0.218656	0.199850

Table 5.2: Inter-document similarity values.

Mainly for practical purposes (envisioning an operational system), we are also interested in the effects that parameters such as the weighting scheme (used for generating document representatives) and the similarity measure (used for computing the similarity between pairs of document representations) have in separating the distribution frequencies for all documents, topical documents, and respectively aspectual documents. In Table 5.2 we show the average similarity between all documents, topical documents and aspectual documents for a combination of independent variables.

The similarity measure and the weighting schemes influence the similarity scale, so comparing absolute values of similarity makes no sense. Instead, we are interested in the effect that these parameters have in distinguishing topics and aspects of topics. If similarities between documents relevant to the same topic, respectively to the same aspect, are significantly higher than similarities between random pairs of documents, then it is likely that a clustering algorithm can more easily identify topics and aspects of topics. We therefore derive in Table 5.3 the ratio of the average similarity between topical documents, respectively aspectual documents, and the average similarity between all documents. Below are the results of the statistical analysis of variance:

Two-way ANOVA: TopicOverAll versus SimilarityMeasure, WeightingScheme

Source	DF	SS	MS	F	P
Similarity	1	0.017035	0.017035	25.92	0.036
Weighting	2	0.161903	0.080952	123.18	0.008 < 0.01 (!!)

		RelFreq	TfIdf	KL
Cosine	Topic/All	2.649599419	2.953178535	3.018271333
	Aspect/All	4.359057426	9.824947875	11.14956944
Dice	Topic/All	2.754686878	3.024254319	3.161809323
	Aspect/All	4.420191723	9.842896113	11.2702029

Table 5.3: Comparison of inter-document similarity values.

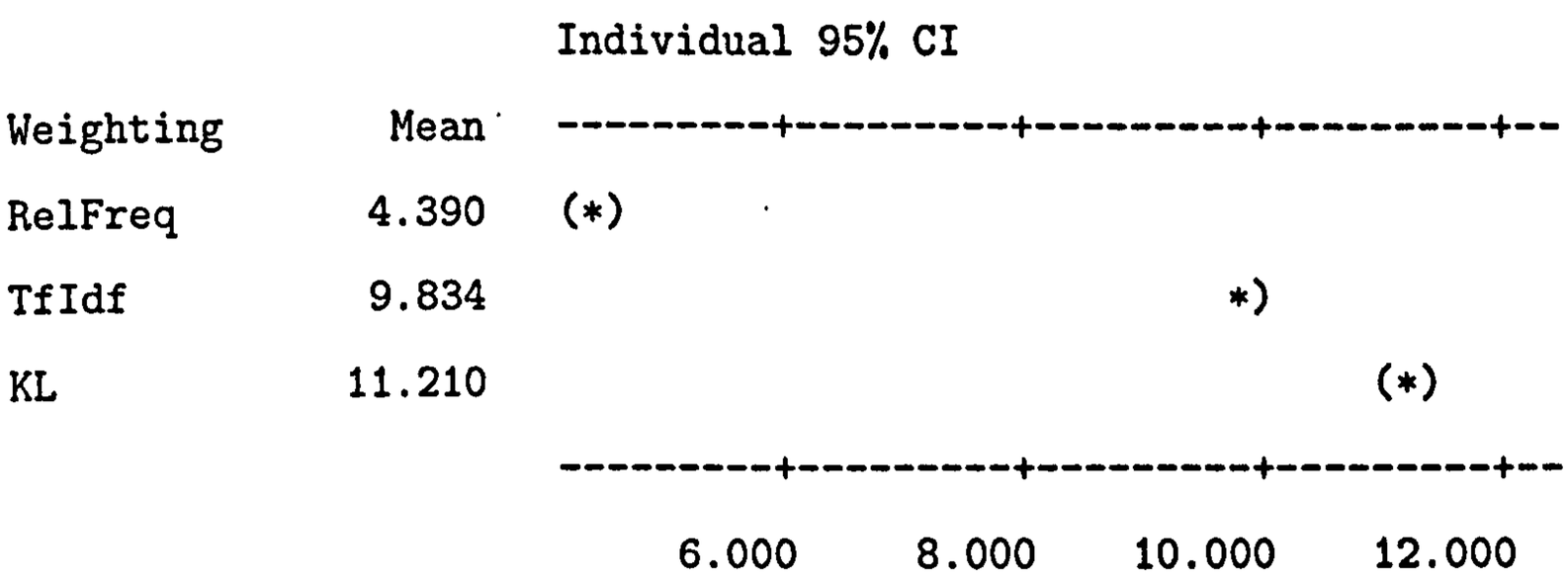
```
                                Individual 95% CI
Similarity          Mean  --+---------+---------+---------+---------
Cosine             2.874  (----------*----------)
Dice               2.980                       (----------*---------)
                          --+---------+---------+---------+---------
                         2.820     2.880     2.940     3.000


                                Individual 95% CI
Weighting           Mean  ------+---------+---------+---------+-----
RelFreq            2.702   (----*----)
TfIdf              2.989                       (----*----)
KL                 3.090                                 (----*----)
                         ------+---------+---------+---------+-----
                            2.700     2.850     3.000     3.150
```

Two-way ANOVA: AspectOverAll versus SimilarityMeasure, WeightingScheme

Source	DF	SS	MS	F	P
Similarity	1	0.00665	0.00665	5.00	0.155
Weighting	2	52.03308	26.01654	2.0E+04	0.000 < 0.01 (!!)

```
                                Individual 95% CI
Similarity          Mean  --------+---------+---------+---------+---
Cosine             8.445  (--------------*--------------)
Dice               8.511                 (--------------*--------------)
                          --------+---------+---------+---------+---
                               8.400     8.460     8.520     8.580
```



The first result of the comparison between the effect of the parameters is that Dice is better (although not significantly better, $p > 0.05$) than Cosine at contributing to the identification of topics and aspects. This is somewhat surprising, as in the context of document clustering Cosine is the most widely used similarity measure. The general consensus is that Cosine is best at grasping topical similarity and that it produces cluster structures that fare well in cluster-based retrieval experiments. Our experiments indicate that Dice is somewhat better at distinguishing topical and aspectual documents. More experiments are needed, with more document collections, to compare a wider range of similarity measures before a final conclusion can be drawn. The experiments here are a beginning.

The second result confirms what was expected, or at least conjectured. The less sophisticated weighting scheme based on the relative frequency fares significantly worse than the other two at identifying both topics and aspects. KL fares somewhat better than TfIdf at identifying topics and significantly better at identifying aspects. The success of our new approach of using the Kullback-Liebler formula for weighting document terms in view of document clustering is especially pleasing as no smoothing or tuning of the formula was done, while in the case of TfIdf the tuned form of the formula as used by Inquery was employed.

Another way to look at this result is from the perspective of the compromise between *representativeness* and *power of discrimination* in document representation. The relative frequency is biased towards representativeness, while the Kullback-Liebler divergence and the tf-idf formula attempt to strike a balance (in the formula $KL = p \cdot \log \frac{p}{q}$, which gives the

		RelFreq	TfIdf	KL
Cosine	All	0.157186	0.088955	0.065393
	Topic	0.553852	0.390477	0.241650
	Aspect	0.543540	0.506575	0.353747
Dice	All	0.127467	0.076278	0.049402
	Topic	0.475461	0.348930	0.241650
	Aspect	0.475461	0.438206	0.353747

Table 5.4: Inter-document similarity values with external weighting scheme.

		RelFreq	TfIdf	KL
Cosine	Topic/All	3.523545354	4.389616288	4.614823514
	Aspect/All	3.457941547	5.694752498	6.903601893
Dice	Topic/All	3.730071313	4.574463343	4.891532073
	Aspect/All	3.654153624	5.744869412	7.160624027

Table 5.5: Comparison of inter-document similarity values with external weighting scheme.

weight of each term in a document, the probability distribution of each term p contributes to representativeness, while the ratio $\frac{p}{q}$, indicates the relative specificity of each term in a document, and therefore it contributes to discriminating a document within a collection). The results seem to indicate that increasing the power of discrimination increases the prominence of topics and aspects, while increasing the representativeness decreases it. This interpretation correlates with Dubin's hypothesis that documents represented with strong discriminators have a greater tendency to cluster than those represented with weak discriminators [Dub96].

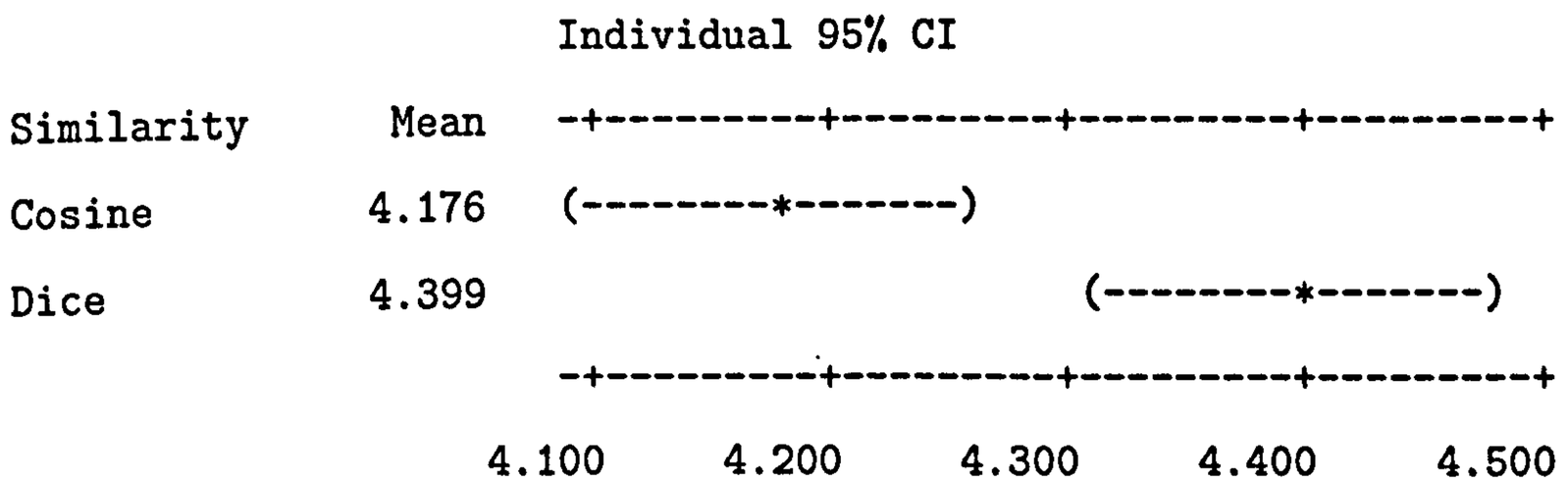
The next step of the experiment looked at the potential of using an **external weighting scheme** for affecting the similarity between the documents and implicitly the clustered structure so that topics and aspects are more prominent. In an operational IR system one can imagine that the system administrator can log the users' queries and compile "Frequently Asked Question" (FAQ). We propose that if these most frequent queries are known, then a weighting scheme based on them should make the topics and aspects covered by them more prominent. For our experiment we used the description of the 6 test topics, extracted the terms in the topic descriptions, and increased the contribution of these terms in the similarity formulae by a factor of 5. The obtained average similarity measures for the same combination of parameters is given in **Table5.5**.

The overall average is, of course, higher with the additional application of the external

weighting scheme. However, the meaningful data is the ratio of the average similarity between topical documents, respectively aspectual documents, and the average similarity between all documents, derived in **Table5.4**. Let us look at the results of the statistical analysis of variance:

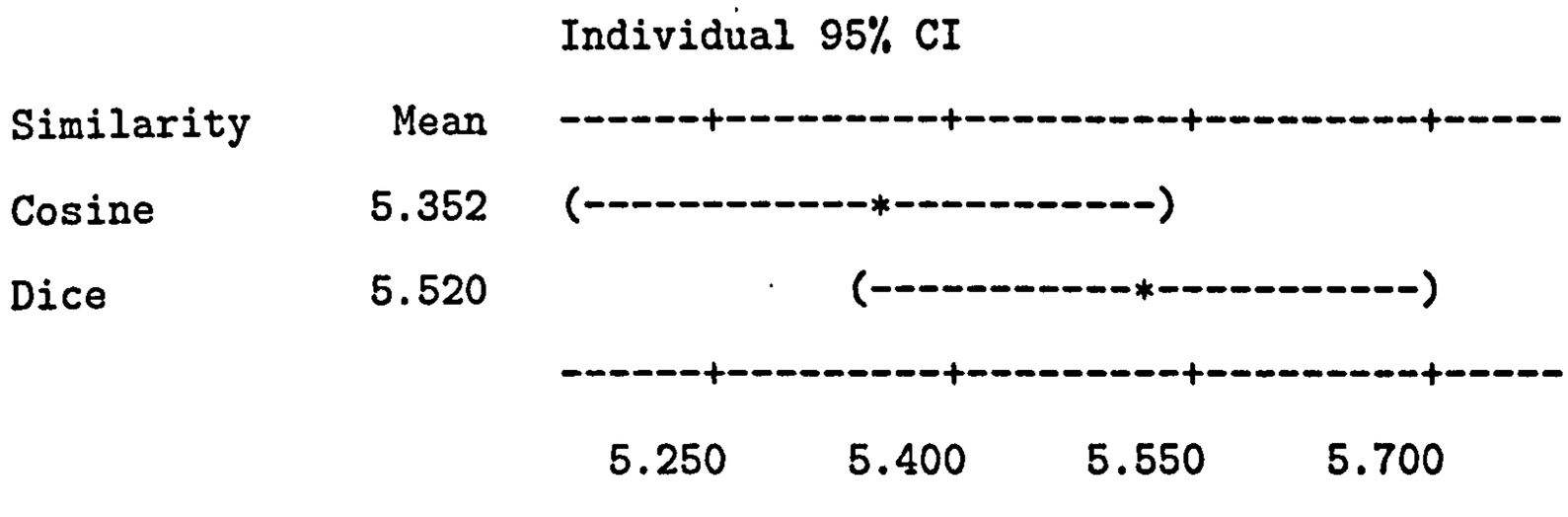
Two-way ANOVA: TopicOverAll versus SimilarityMeasure, WeightingScheme

Source	DF	SS	MS	F	P
Similarity	1	0.07439	0.07439	64.53	0.015 < 0.05 (!)
Weighting	2	1.38243	0.69121	599.57	0.002 < 0.01 (!!)



Two-way ANOVA: AspectOverAll versus SimilarityMeasure, WeightingScheme

Source	DF	SS	MS	F	P
Similarity	1	0.04223	0.04223	7.47	0.112
Weighting	2	12.32469	6.16235	1089.85	0.001 < 0.01 (!!)

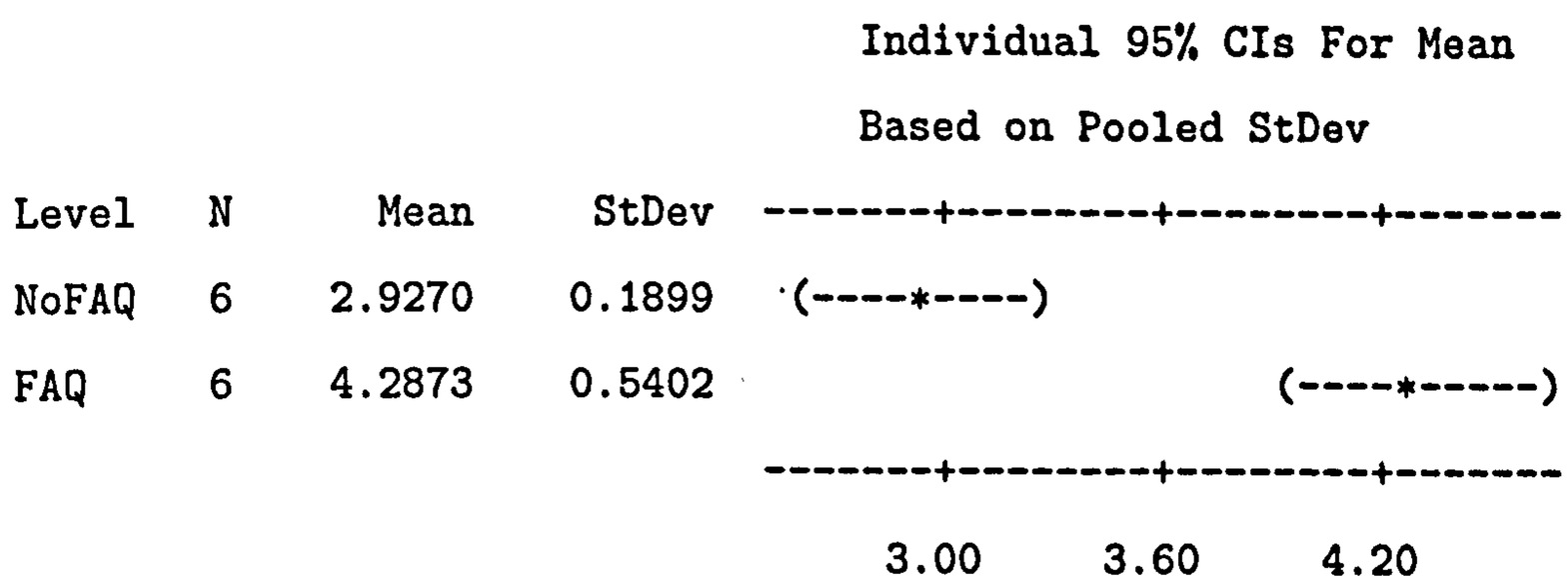


It is apparent that the same kind of conclusions can be drawn as for the case when no external weighting was used, but now some conclusions are stronger: Dice is significantly better than Cosine at identifying topics and KL is significantly better than TfIdf at identifying both topics and aspects.

When comparing the two cases, without and with external weighting, the data in the tables suggests that, as expected, the topics are more visible in the latter case It may seem surprising that the aspects are less visible. However, only rather general descriptions of the topics were available as *FAQ*. There were no descriptions of the aspects. In fact, the aspects were established by the TREC experiment judges post-experiment, based on the documents proposed as relevant by the Interactive TREC-8 experiment participants. Therefore, the aspects are blurred by applying topic-specific weighting. The statistical analysis confirms this interpretation of the data:

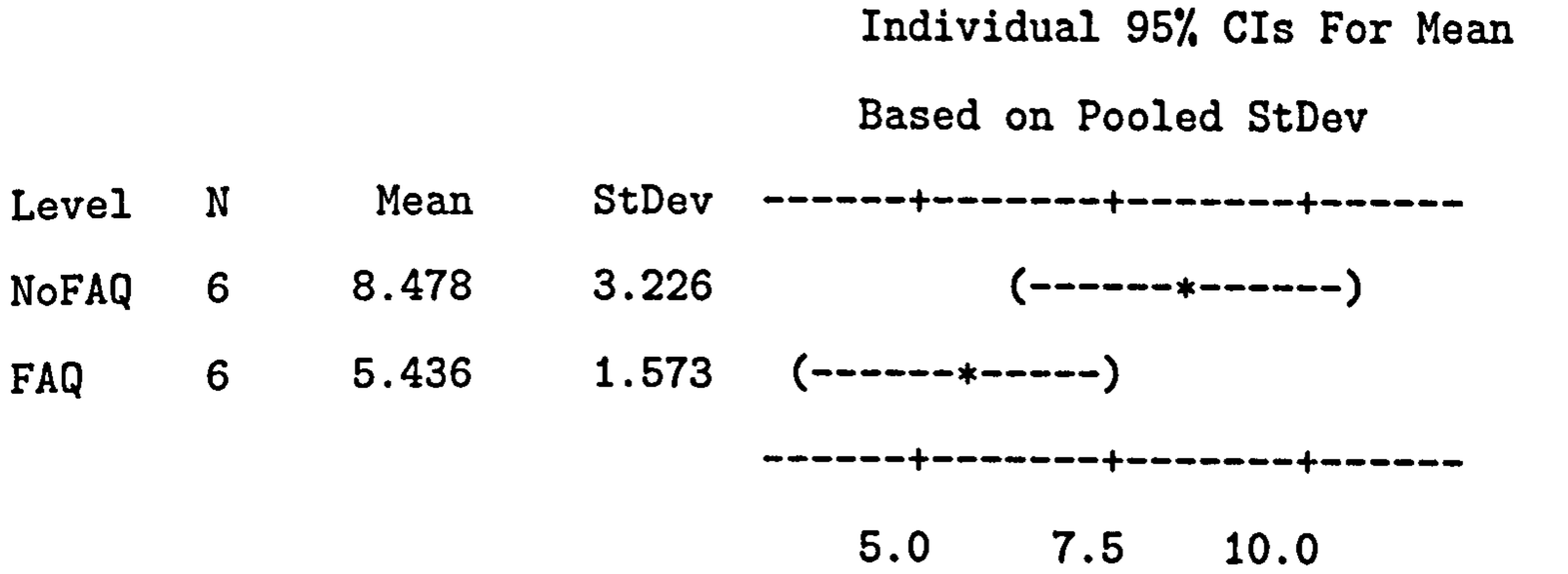
One-way ANOVA: TopicOverAll versus ExternalWeight

Source	DF	SS	MS	F	P
External	1	5.552	5.552	33.87	0.000 < 0.01 (!!)



One-way ANOVA: AspectOverAll versus ExternalWeight

Source	DF	SS	MS	F	P
External	1	27.76	27.76	4.31	0.065
Error	10	64.42	6.44		
Total	11	92.18			



So, the external, *FAQ*-based weighting scheme increases the topic visibility highly significantly, but decreases (although not significantly) the aspect visibility. It is probably the case that, if descriptions of typical aspects were available, the visibility of the aspects could also be increased by applying an appropriate external weighting scheme.

5.3 Topic distribution over the cluster structure

5.3.1 Approach

While the previous section looked at how well topics can be identified by examining the inter-document similarities, this section looks at how well clustering algorithms recognize these topics and separate them into clusters. We are particularly interested in *complete-link* (CL) and *group-average* (GA), the two hierarchic clustering algorithms identified as

candidates for clustering source collection for the purpose of guiding exploration of a domain of interest and, therefore, for supporting mediation.

In 30 years of research in using document clustering for Information Retrieval, experimental results have often proved inconsistent or even contradictory between collections. Therefore, the results obtained here may not be safely generalised. This does not invalidate the idea of mediated access, it just imposes a methodological restriction: before applying clustering on a collection used as source collection for mediation, it is safer to experiment with a set of clustering parameters in order to obtain the best structure.

As a starting point for clustering we are using the similarity matrices computed as part of the experiment described in the previous section. For the sake of clarity, we will first use the similarity matrix built with one combination of similarity measure and weighting scheme, and explore in some detail the cluster structures obtained with complete-link and group-average. We will devise a measure of cluster quality and, if we consider it adequate, will look at the effect on cluster quality of the other independent variables such as similarity measure and weighting scheme. He have chosen for this analysis the cluster structures built with the Cosine similarity measure (the most commonly used) and Kullback-Liebler weighting (not used before as far as we know).

Apart from testing the *cluster hypothesis* through tests that looked at the separation between relevant and non-relevant documents (for each topic of a test collection), traditional clustering experiments also looked at the quality of the cluster structure. However, even the practical, retrieval-oriented tests that evaluated the capacity of the cluster structure to support effective document retrieval were limited to the batch-retrieval scenario and to computing the effectiveness of cluster-based retrieval. The methodologies used in those experiments are clearly inadequate for indicating the quality of clustering in an interactive, task-oriented setting, as is the case for the mediation scenario. Specific methodologies need to be developed for the new conditions, so that they take into account the use of the cluster structure.

In the case of WebCluster the hierarchic structure of a collection, obtained through

clustering, is used for exploring the topics of a domain of interest and for identifying clusters that are representative for a certain information need. In our context the clustering is good if documents relevant for a certain topic are grouped in a relatively small number of clusters so that, with the help of search tools, the exploration can be limited to a small subdomain. While such tests are relatively straightforward in the case of non-hierarchic clustering [HP96], a hierarchic structure imposes the problem of *granularity*: depending on the level of the hierarchy, some documents may or may not be viewed as belonging to the same cluster. The size of the clusters is also a problem: it may be more convenient for a user to have to explore say three small clusters than just one large cluster. Therefore, the size of the clusters and their level of specialization (or level of similarity) needs to be taken into account when evaluating how many clusters a user needs to explore in order to cover a topic.

Algorithm 5 Evaluation of top clusters in the cluster hierarchy.

```
Cut the hierarchic structure at top level.
Obtain a partitioning into clusters C_1, ..., C_k.
for all topic T_i do
  for all cluster C_j do
    Compute recall R_{i,j} = percentage of documents relevant to T_i that are in C_j.
    Compute precision P_{i,j} = percentage of documents in C_j that are relevant to T_i.
  end for
  Sort clusters C_1, ..., C_k based on their recall.
end for
```

We can get a rough idea of the distribution of the documents covering various topics by looking at how well the top clusters in the hierarchy cover each of the topics in the test collection. **Algorithm 5** cuts the structure at top level and, for each topic, ranks the obtained clusters based on the number of relevant documents they contained (and therefore on recall).

5.3.2 Experimental results

Table 5.6 shows the result of **Algorithm 5** on the hierarchy obtained by applying a group-average algorithm to the source collection of 747 documents, for which the similarity matrix had been calculated based on the Cosine similarity measure and the Kullback-Liebler weighting scheme. For each of the 6 topics, separated into columns, the 48 top

Topic 1 : 408			Topic 2 : 414			Topic 3 : 428			Topic 4 : 431			Topic 5 : 438			Topic 6 : 446		
Docs :	Rel	R	Docs :	Rel	R	Docs :	Rel	R	Docs :	Rel	R	Docs :	Rel	R	Docs :	Rel	R
73 :	18	0.51	79 :	8	1.00	49 :	12	0.60	21 :	10	0.30	97 :	22	0.42	98 :	20	0.69
40 :	13	0.37	2 :	0	0	16 :	2	0.10	53 :	7	0.21	19 :	5	0.10	73 :	5	0.17
19 :	2	0.06	5 :	0	0	97 :	2	0.10	5 :	3	0.09	11 :	3	0.06	5 :	1	0.03
24 :	1	0.03	6 :	0	0	5 :	1	0.05	6 :	2	0.06	98 :	3	0.06	11 :	1	0.03
97 :	1	0.03	3 :	0	0	6 :	1	0.05	2 :	1	0.03	2 :	2	0.04	12 :	1	0.03
2 :	0	0	1 :	0	0	19 :	1	0.05	3 :	1	0.03	7 :	2	0.04	97 :	1	0.03
5 :	0	0	6 :	0	0	98 :	1	0.05	3 :	1	0.03	11 :	2	0.04	2 :	0	0
6 :	0	0	2 :	0	0	2 :	0	0	3 :	1	0.03	12 :	2	0.04	5 :	0	0
3 :	0	0	11 :	0	0	5 :	0	0	5 :	1	0.03	16 :	2	0.04	6 :	0	0
1 :	0	0	2 :	0	0	6 :	0	0	5 :	1	0.03	73 :	2	0.04	3 :	0	0
79 :	0	0	7 :	0	0	3 :	0	0	5 :	1	0.03	2 :	1	0.02	1 :	0	0
6 :	0	0	19 :	0	0	1 :	0	0	5 :	1	0.03	5 :	1	0.02	70 :	0	0
2 :	0	0	16 :	0	0	79 :	0	0	5 :	1	0.03	6 :	1	0.02	6 :	0	0
11 :	0	0	2 :	0	0	6 :	0	0	6 :	1	0.03	8 :	1	0.02	2 :	0	0
2 :	0	0	97 :	0	0	2 :	0	0	6 :	1	0.03	11 :	1	0.02	2 :	0	0
7 :	0	0	6 :	0	0	11 :	0	0	2 :	0	0	24 :	1	0.02	7 :	0	0
16 :	0	0	49 :	0	0	2 :	0	0	5 :	0	0	79 :	1	0.02	19 :	0	0
2 :	0	0	8 :	0	0	7 :	0	0	6 :	0	0	2 :	0	0	16 :	0	0
6 :	0	0	40 :	0	0	2 :	0	0	3 :	0	0	5 :	0	0	2 :	0	0
49 :	0	0	73 :	0	0	8 :	0	0	1 :	0	0	6 :	0	0	6 :	0	0
8 :	0	0	24 :	0	0	40 :	0	0	79 :	0	0	3 :	0	0	49 :	0	0
3 :	0	0	3 :	0	0	73 :	0	0	2 :	0	0	1 :	0	0	8 :	0	0
5 :	0	0	5 :	0	0	24 :	0	0	11 :	0	0	6 :	0	0	40 :	0	0
98 :	0	0	98 :	0	0	3 :	0	0	2 :	0	0	2 :	0	0	24 :	0	0
11 :	0	0	11 :	0	0	11 :	0	0	7 :	0	0	49 :	0	0	3 :	0	0
2 :	0	0	2 :	0	0	2 :	0	0	19 :	0	0	40 :	0	0	5 :	0	0
12 :	0	0	12 :	0	0	12 :	0	0	16 :	0	0	3 :	0	0	11 :	0	0
11 :	0	0	11 :	0	0	11 :	0	0	2 :	0	0	5 :	0	0	2 :	0	0
5 :	0	0	5 :	0	0	5 :	0	0	97 :	0	0	2 :	0	0	11 :	0	0
6 :	0	0	6 :	0	0	6 :	0	0	49 :	0	0	6 :	0	0	5 :	0	0
4 :	0	0	4 :	0	0	4 :	0	0	8 :	0	0	4 :	0	0	6 :	0	0
5 :	0	0	5 :	0	0	5 :	0	0	40 :	0	0	5 :	0	0	4 :	0	0
2 :	0	0	2 :	0	0	2 :	0	0	73 :	0	0	2 :	0	0	2 :	0	0
5 :	0	0	5 :	0	0	5 :	0	0	24 :	0	0	5 :	0	0	5 :	0	0
3 :	0	0	3 :	0	0	3 :	0	0	3 :	0	0	3 :	0	0	3 :	0	0
3 :	0	0	3 :	0	0	3 :	0	0	5 :	0	0	3 :	0	0	3 :	0	0
21 :	0	0	21 :	0	0	21 :	0	0	98 :	0	0	21 :	0	0	21 :	0	0
2 :	0	0	2 :	0	0	2 :	0	0	11 :	0	0	2 :	0	0	2 :	0	0
3 :	0	0	3 :	0	0	3 :	0	0	2 :	0	0	3 :	0	0	3 :	0	0
53 :	0	0	53 :	0	0	53 :	0	0	12 :	0	0	53 :	0	0	53 :	0	0
5 :	0	0	5 :	0	0	5 :	0	0	11 :	0	0	5 :	0	0	5 :	0	0
5 :	0	0	5 :	0	0	5 :	0	0	5 :	0	0	5 :	0	0	5 :	0	0
5 :	0	0	5 :	0	0	5 :	0	0	6 :	0	0	5 :	0	0	5 :	0	0
3 :	0	0	3 :	0	0	3 :	0	0	4 :	0	0	3 :	0	0	3 :	0	0
4 :	0	0	4 :	0	0	4 :	0	0	2 :	0	0	4 :	0	0	4 :	0	0
5 :	0	0	5 :	0	0	5 :	0	0	3 :	0	0	5 :	0	0	5 :	0	0
6 :	0	0	6 :	0	0	6 :	0	0	4 :	0	0	6 :	0	0	6 :	0	0
2 :	0	0	2 :	0	0	2 :	0	0	2 :	0	0	2 :	0	0	2 :	0	0

Table 5.6: The distribution of relevant documents over the top clusters of a structure obtained with Group-Average clustering. Parameters: Cosine, KL.

clusters are ranked based on the number of relevant documents. For each top cluster, the table shows the number of documents in the cluster (*Docs*), the number of relevant documents for the topic (*Rel*), as well as the recall of the cluster (*R*). The precision of each cluster, which is not essential for this experiment, was not included in the table for reasons of space, but can be estimated by the reader ($P = Rel/Docs$). It is apparent from the table that:

- for each topic, all the relevant documents are grouped in a relatively small number of clusters.
- for each topic, the bulk of the relevant documents are grouped in 1, 2 or 3 clusters.
- the topics are well separated by the clustering process.

The grouping of relevant documents into a small number of 'good' clusters is significantly different from what is obtained by randomly allocating the relevant documents to the top clusters and ranking the clusters based on the number of relevant documents. **Figure 5.5** shows the difference. It is apparent that, on average over the topics, the best top cluster contains approximately 58% of the relevant documents and that even if the user wanted all the relevant documents, only a small number of documents would need to be explored.

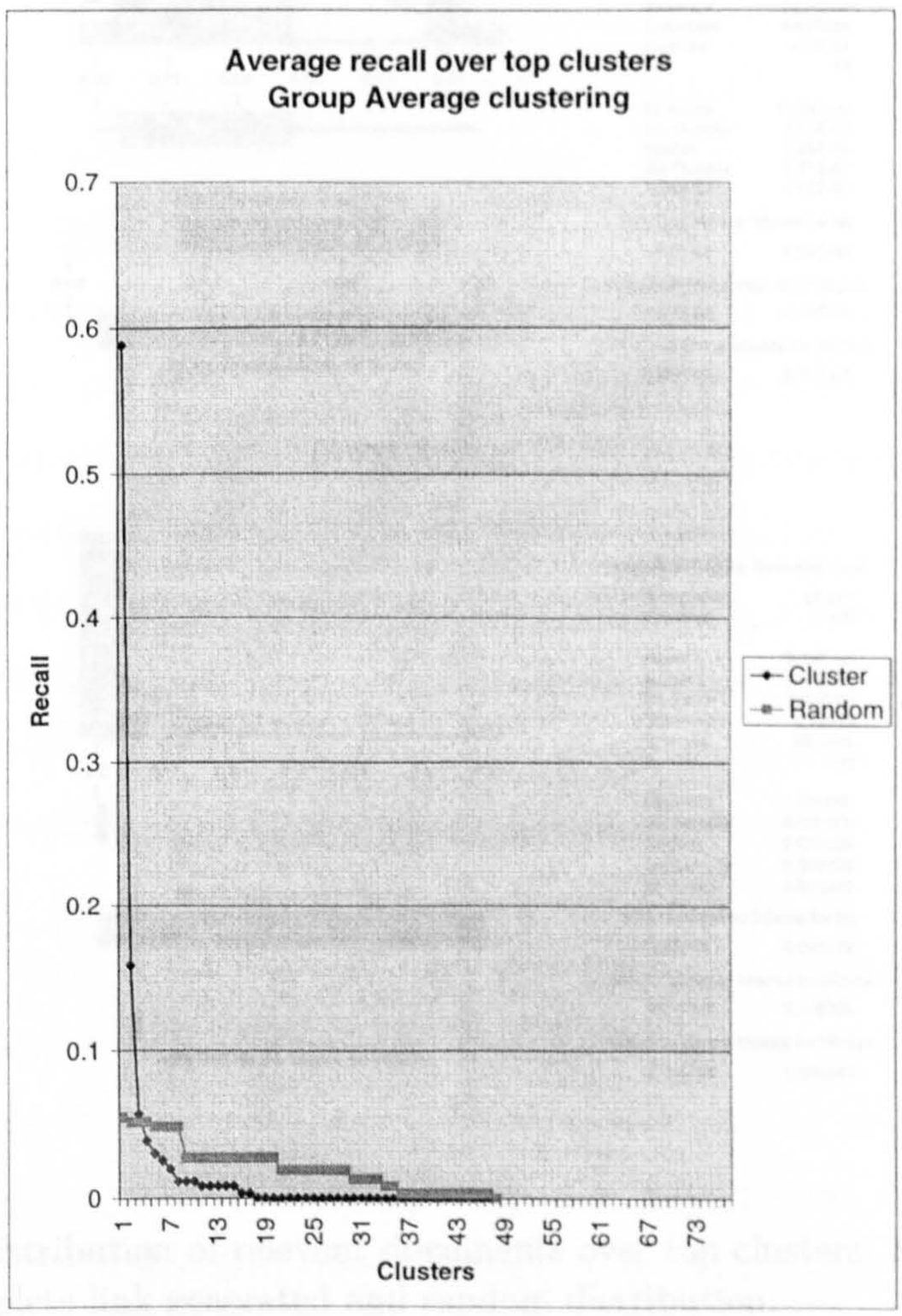


Figure 5.5: The distribution of recall over top clusters. Compare group-average with random distribution.

A statistical comparison between the grouping of relevant documents in top clusters produced by the group-average clustering algorithm and a random distribution of these documents is presented in **Figure 5.6**. The differences in skewness (given by the position

of the frequency distribution peak) and in kurtosis (given by the sharpness of the peak) are highly significant, with the clustering algorithm successfully grouping the relevant documents in a small number of top clusters.

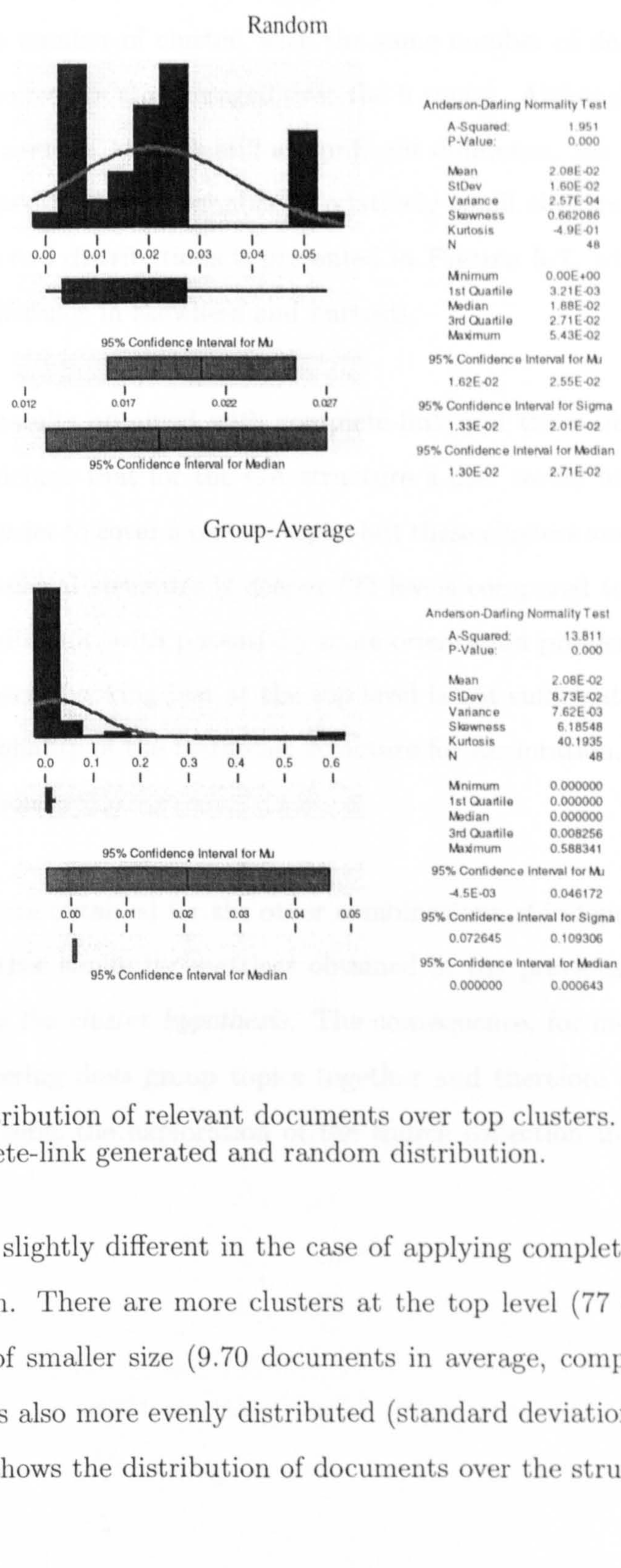


Figure 5.6: The distribution of relevant documents over top clusters. Statistical comparison between complete-link generated and random distribution.

The situation is slightly different in the case of applying complete-link for clustering the source collection. There are more clusters at the top level (77 compared to 48 for GA) and they are of smaller size (9.70 documents in average, compared to 15.56); the size of the clusters is also more evenly distributed (standard deviation 7.48, compared to 24.64). **Table 5.7** shows the distribution of documents over the structure obtained with

complete-link clustering. Again, the clusters are ranked based on the number of relevant documents for each of the 6 topics.

Figure 5.7 compares the result of dispersion of the relevant documents over the top clusters obtained by complete-link clustering with the dispersion of the relevant documents over the same number of cluster, with the same number of documents, through a random process. The results are averaged over the 6 topics. Although not as striking as in the case of group-average, there is still a significant difference: the bulk of the relevant documents are grouped in a small number of relatively small clusters. A statistical comparison between the two distributions is presented in **Figure 5.7**, which also indicates a highly significant difference in skewness and kurtosis.

Comparing the results obtained with complete-link and those obtained with group-average, one can conclude that for the GA structure a user would have to explore fewer top level clusters in order to cover a certain topic, but these clusters are significantly larger. Moreover, the hierarchical structure is deeper (22 levels compared to 14 for CL), so the exploration is more difficult, with potentially more orientation problems and heavier cognitive load for the user. Looking just at the top level is not sufficient to draw conclusion with regards to the quality of the hierarchic structure for exploration. We devise a better test in the next section.

Similar results were obtained for the other combinations of independent variables, respectively for the other similarity matrices obtained in the previous experiment which, once again, *confirms the cluster hypothesis*. The consequence, for mediation, is the confirmation that clustering does group topics together and therefore a user investigating a certain topic can limit the exploration of the source collection to a small number of clusters.

Topic 1 : 408			Topic 2 : 414			Topic 3 : 428			Topic 4 : 431			Topic 5 : 438			Topic 6 : 446		
Docs :	Rel	R	Docs :	Rel	R	Docs :	Rel	R	Docs :	Rel	R	Docs :	Rel	R	Docs :	Rel	R
15 :	10	0.29	41 :	5	0.63	20 :	7	0.35	9 :	5	0.15	42 :	14	0.27	16 :	5	0.17
6 :	5	0.14	31 :	2	0.25	11 :	2	0.10	15 :	4	0.12	17 :	5	0.10	19 :	5	0.17
19 :	5	0.14	5 :	1	0.13	13 :	2	0.10	17 :	4	0.12	11 :	4	0.08	7 :	3	0.10
7 :	3	0.09	6 :	0	0	14 :	2	0.10	5 :	3	0.09	10 :	3	0.06	5 :	2	0.07
6 :	2	0.06	5 :	0	0	3 :	1	0.05	18 :	3	0.09	3 :	2	0.04	7 :	2	0.07
8 :	2	0.06	6 :	0	0	4 :	1	0.05	6 :	2	0.06	8 :	2	0.04	8 :	2	0.07
8 :	2	0.06	6 :	0	0	5 :	1	0.05	8 :	2	0.06	9 :	2	0.04	3 :	1	0.03
11 :	2	0.06	10 :	0	0	10 :	1	0.05	10 :	2	0.06	11 :	2	0.04	6 :	1	0.03
6 :	1	0.03	9 :	0	0	10 :	1	0.05	2 :	1	0.03	11 :	2	0.04	6 :	1	0.03
13 :	1	0.03	7 :	0	0	10 :	1	0.05	3 :	1	0.03	13 :	2	0.04	7 :	1	0.03
15 :	1	0.03	10 :	0	0	13 :	1	0.05	4 :	1	0.03	4 :	1	0.02	8 :	1	0.03
17 :	1	0.03	7 :	0	0	6 :	0	0	7 :	1	0.03	4 :	1	0.02	10 :	1	0.03
5 :	0	0	6 :	0	0	5 :	0	0	8 :	1	0.03	4 :	1	0.02	10 :	1	0.03
6 :	0	0	3 :	0	0	6 :	0	0	10 :	1	0.03	4 :	1	0.02	10 :	1	0.03
10 :	0	0	7 :	0	0	6 :	0	0	14 :	1	0.03	5 :	1	0.02	11 :	1	0.03
9 :	0	0	6 :	0	0	9 :	0	0	19 :	1	0.03	6 :	1	0.02	11 :	1	0.03
10 :	0	0	10 :	0	0	7 :	0	0	6 :	0	0	6 :	1	0.02	6 :	0	0
7 :	0	0	6 :	0	0	10 :	0	0	5 :	0	0	7 :	1	0.02	5 :	0	0
3 :	0	0	11 :	0	0	7 :	0	0	6 :	0	0	7 :	1	0.02	6 :	0	0
31 :	0	0	8 :	0	0	6 :	0	0	6 :	0	0	7 :	1	0.02	10 :	0	0
41 :	0	0	42 :	0	0	3 :	0	0	10 :	0	0	13 :	1	0.02	9 :	0	0
7 :	0	0	13 :	0	0	31 :	0	0	9 :	0	0	14 :	1	0.02	7 :	0	0
6 :	0	0	5 :	0	0	41 :	0	0	7 :	0	0	16 :	1	0.02	6 :	0	0
10 :	0	0	9 :	0	0	7 :	0	0	10 :	0	0	19 :	1	0.02	3 :	0	0
6 :	0	0	8 :	0	0	6 :	0	0	7 :	0	0	6 :	0	0	31 :	0	0
11 :	0	0	2 :	0	0	6 :	0	0	6 :	0	0	5 :	0	0	41 :	0	0
8 :	0	0	6 :	0	0	11 :	0	0	3 :	0	0	6 :	0	0	6 :	0	0
42 :	0	0	2 :	0	0	8 :	0	0	31 :	0	0	10 :	0	0	10 :	0	0
13 :	0	0	19 :	0	0	42 :	0	0	41 :	0	0	9 :	0	0	6 :	0	0
5 :	0	0	17 :	0	0	5 :	0	0	7 :	0	0	10 :	0	0	8 :	0	0
9 :	0	0	6 :	0	0	9 :	0	0	6 :	0	0	7 :	0	0	42 :	0	0
8 :	0	0	4 :	0	0	8 :	0	0	6 :	0	0	6 :	0	0	13 :	0	0
2 :	0	0	4 :	0	0	2 :	0	0	11 :	0	0	3 :	0	0	5 :	0	0
6 :	0	0	2 :	0	0	6 :	0	0	42 :	0	0	31 :	0	0	9 :	0	0
2 :	0	0	10 :	0	0	2 :	0	0	13 :	0	0	41 :	0	0	2 :	0	0
6 :	0	0	19 :	0	0	19 :	0	0	8 :	0	0	7 :	0	0	2 :	0	0
4 :	0	0	11 :	0	0	17 :	0	0	2 :	0	0	6 :	0	0	19 :	0	0
4 :	0	0	5 :	0	0	6 :	0	0	6 :	0	0	10 :	0	0	17 :	0	0
2 :	0	0	8 :	0	0	4 :	0	0	2 :	0	0	6 :	0	0	6 :	0	0
10 :	0	0	7 :	0	0	2 :	0	0	19 :	0	0	11 :	0	0	4 :	0	0
19 :	0	0	5 :	0	0	10 :	0	0	17 :	0	0	8 :	0	0	4 :	0	0
11 :	0	0	8 :	0	0	19 :	0	0	6 :	0	0	5 :	0	0	2 :	0	0
5 :	0	0	8 :	0	0	11 :	0	0	4 :	0	0	9 :	0	0	8 :	0	0
8 :	0	0	3 :	0	0	5 :	0	0	4 :	0	0	8 :	0	0	5 :	0	0
7 :	0	0	4 :	0	0	8 :	0	0	2 :	0	0	2 :	0	0	8 :	0	0
5 :	0	0	14 :	0	0	7 :	0	0	10 :	0	0	2 :	0	0	3 :	0	0
3 :	0	0	8 :	0	0	5 :	0	0	19 :	0	0	19 :	0	0	4 :	0	0
4 :	0	0	10 :	0	0	8 :	0	0	11 :	0	0	17 :	0	0	14 :	0	0
14 :	0	0	17 :	0	0	8 :	0	0	5 :	0	0	6 :	0	0	8 :	0	0
8 :	0	0	18 :	0	0	3 :	0	0	8 :	0	0	2 :	0	0	10 :	0	0
10 :	0	0	8 :	0	0	4 :	0	0	7 :	0	0	10 :	0	0	17 :	0	0
17 :	0	0	16 :	0	0	8 :	0	0	5 :	0	0	5 :	0	0	18 :	0	0
18 :	0	0	17 :	0	0	10 :	0	0	8 :	0	0	8 :	0	0	8 :	0	0
8 :	0	0	4 :	0	0	17 :	0	0	8 :	0	0	8 :	0	0	17 :	0	0
16 :	0	0	8 :	0	0	18 :	0	0	3 :	0	0	8 :	0	0	4 :	0	0
17 :	0	0	10 :	0	0	8 :	0	0	4 :	0	0	10 :	0	0	8 :	0	0
4 :	0	0	11 :	0	0	16 :	0	0	8 :	0	0	17 :	0	0	10 :	0	0
8 :	0	0	11 :	0	0	17 :	0	0	16 :	0	0	18 :	0	0	11 :	0	0
10 :	0	0	6 :	0	0	4 :	0	0	17 :	0	0	8 :	0	0	11 :	0	0
11 :	0	0	6 :	0	0	8 :	0	0	4 :	0	0	8 :	0	0	6 :	0	0
6 :	0	0	3 :	0	0	11 :	0	0	8 :	0	0	10 :	0	0	6 :	0	0
6 :	0	0	9 :	0	0	11 :	0	0	10 :	0	0	11 :	0	0	3 :	0	0
3 :	0	0	3 :	0	0	6 :	0	0	11 :	0	0	6 :	0	0	9 :	0	0
9 :	0	0	15 :	0	0	6 :	0	0	11 :	0	0	6 :	0	0	15 :	0	0
3 :	0	0	19 :	0	0	3 :	0	0	6 :	0	0	3 :	0	0	19 :	0	0
15 :	0	0	4 :	0	0	9 :	0	0	3 :	0	0	3 :	0	0	4 :	0	0
19 :	0	0	2 :	0	0	3 :	0	0	9 :	0	0	15 :	0	0	2 :	0	0
4 :	0	0	7 :	0	0	15 :	0	0	10 :	0	0	19 :	0	0	7 :	0	0
2 :	0	0	10 :	0	0	19 :	0	0	13 :	0	0	4 :	0	0	13 :	0	0
7 :	0	0	13 :	0	0	4 :	0	0	20 :	0	0	2 :	0	0	20 :	0	0
10 :	0	0	20 :	0	0	2 :	0	0	4 :	0	0	20 :	0	0	4 :	0	0
20 :	0	0	4 :	0	0	7 :	0	0	11 :	0	0	4 :	0	0	11 :	0	0
4 :	0	0	11 :	0	0	10 :	0	0	15 :	0	0	15 :	0	0	15 :	0	0
11 :	0	0	15 :	0	0	4 :	0	0	15 :	0	0	15 :	0	0	15 :	0	0
5 :	0	0	15 :	0	0	15 :	0	0	5 :	0	0	5 :	0	0	5 :	0	0
3 :	0	0	3 :	0	0	15 :	0	0	3 :	0	0	3 :	0	0	3 :	0	0
3 :	0	0	3 :	0	0	3 :	0	0	3 :	0	0	3 :	0	0	3 :	0	0

Table 5.7: The distribution of relevant documents over the top clusters of a structure obtained with Complete-Link clustering. Parameters: Cosine, KL.

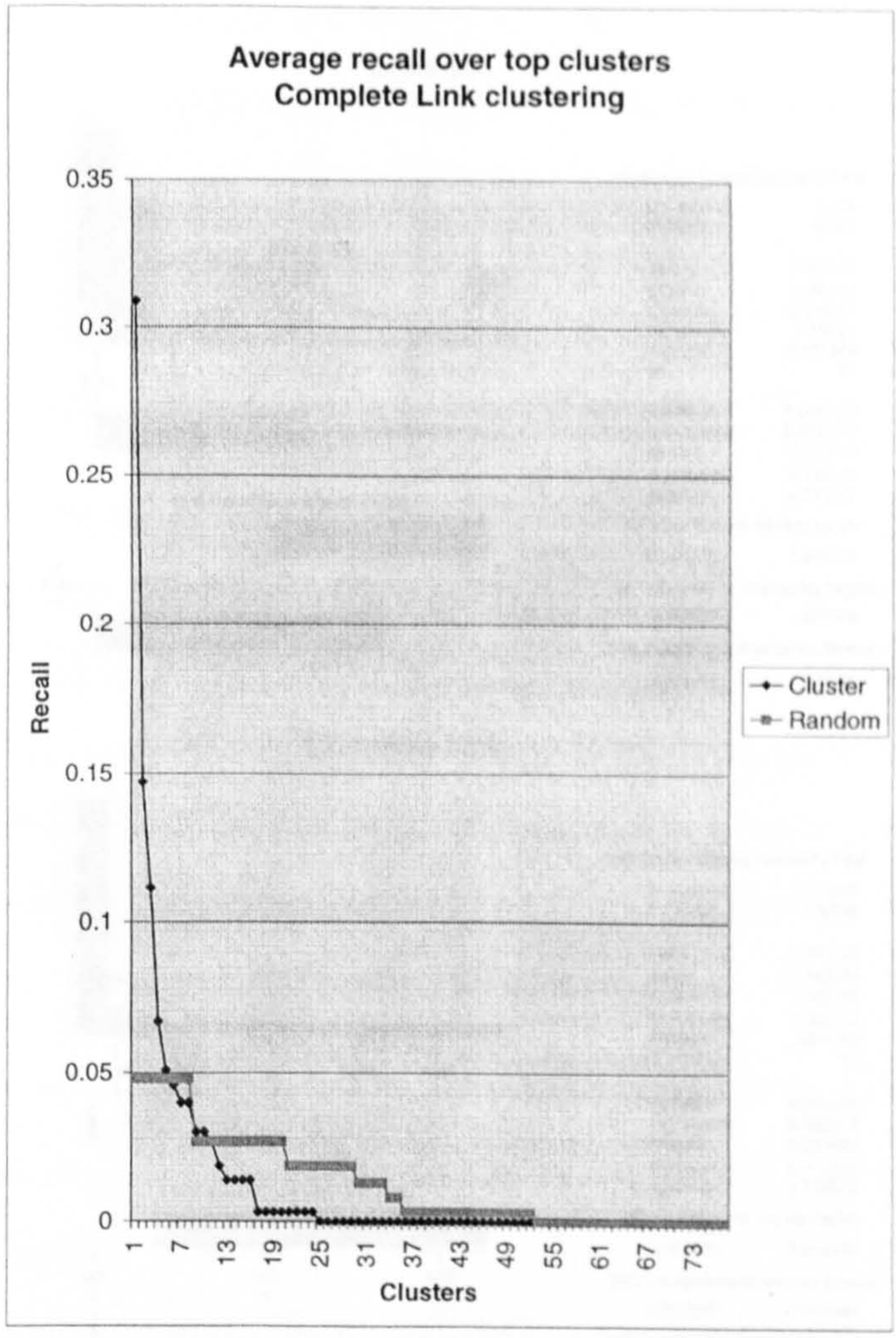


Figure 5.7: The distribution of recall over top clusters. Compare complete-link with random distribution.

5.4 Topic distribution in more detail

5.4.1 The approach

The successful experiment described in the previous section was significant from the point of view of testing the cluster hypothesis. However, as it only looks at the top level of the cluster hierarchy, it does not fully convey the distribution of the topics over the structure and it offers insufficient evidence as to the potential of clustering for mediation. Moreover, due to the difference in cluster sizes produced by the two clustering algorithms, complete-link and group-average, it is difficult to assess which structure is better for supporting the user's exploration. The number of top level clusters that need to be looked at is not a sufficient indication of the user's effort, due to the difference in size and level of inter-

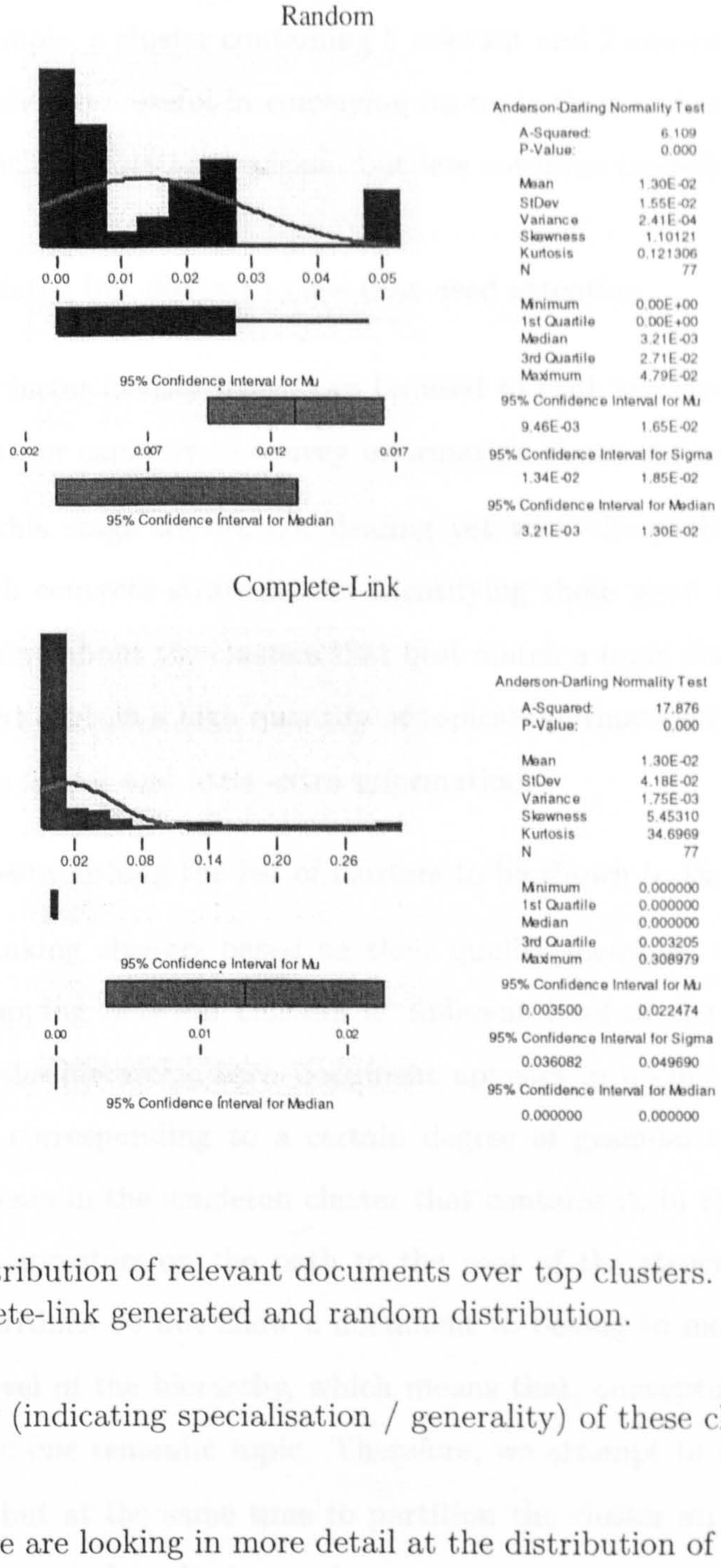


Figure 5.8: The distribution of relevant documents over top clusters. Statistical comparison between complete-link generated and random distribution.

document similarity (indicating specialisation / generality) of these clusters.

In this section we are looking in more detail at the distribution of relevant documents over the hierarchic structure and are trying to assign some measure of quality to this distribution. For the purpose of exploration of the cluster structure, as the first stage in mediation, a 'good' structure should contain 'good' clusters. A good cluster is, intuitively, a cluster that contains as many documents that are representative for a certain topic as possible, so that the user can get an idea of the topic, and as few extra documents as possible, so that the user's idea of the topic is not diluted. We argue that for exploratory

searches a decrease in precision should be accepted for the sake of capturing the topics of the domain. For example, a cluster containing 5 relevant and 2 non-relevant (but related) documents is probably more useful in conveying its topic than a cluster with just 2 relevant documents, which has better precision, but less coverage (recall).

There are two related but distinct issues that need attention:

1. A measure of cluster quality which can be used to rank the clusters based on their *informativeness*, or capacity to convey information about the topic of interest.

   Note that at this stage we are not dealing yet with the problem of representing clusters or with concrete strategies for identifying these good clusters. Therefore, we are not talking about the clusters that best match a topic description, but about the clusters that contain a high quantity of topical information (as judged by expert relevance judgements) and little extra information.

2. A method for establishing the list of clusters to be shown to the user.

   Apart from ranking clusters based on their quality measure, we need to consider possible overlapping between clusters at different level of the hierarchy. Due to the nature of the hierarchy, each document appears in a cluster at every level of the hierarchy, corresponding to a certain degree of granularity. More exactly, a document appears in the singleton cluster that contains it, in that cluster's parent, and in all the ancestors on the path to the root of the structure. However, our clustering algorithms do not allow a document to belong to more than one cluster at a certain level of the hierarchy, which means that, conceptually, one document only belongs to one semantic topic. Therefore, we attempt to rank the clusters of the collection but at the same time to partition the cluster structure so that each document is presented to the user only once.

While we dedicate the full next section to the former of these problems, which is a difficult and controversial one, **Algorithm 6** describes a simple and intuitive procedure that offers a solution to the latter. Its output is a list of clusters ranked in decreasing order of their quality which also creates a partition of the document collection, to the effect that each document only appears in one of the clusters. The clusters whose quality

Algorithm 6 Best clusters partitioning of the hierarchic structure.

```
for all topic T_i do
    Compute quality Q for each cluster in the hierarchy.
    Rank clusters based on Q, creating the list ClusterList.
    Visit the clusters in ClusterList in order.
    for all cluster C do
        Eliminate the ancestors and descendants of C from the rest of the list.
    end for
    The remaining ClusterList forms a ranked partition of the collection.
end for
```

is estimated to be 0 (not containing any relevant documents) are not ranked or presented to the user. Conceptually they form the 'garbage' cluster of the partition.

5.4.2 Cluster quality

This section discusses possible measures of cluster quality based on which ranked clusters should be presented to the user.

Let us concentrate on what constitutes a good cluster for a user exploring a document collection in an unfamiliar domain. *Recall* is desirable, so that the user gets good coverage of the topic of interest. However, if the cluster is very large and has low precision, then the user needs to make some effort to identify relevant documents and may get confused by the non-relevant documents. At the limit, the root cluster has 100% recall, but exploring it (i.e. the full collection) would mean not making use of the searching capabilities of the system. On the other hand, *precision* is desirable, so that the user finds relevant documents easily and is not confused by non-relevant documents. However, if the cluster is very small and has low recall, then the user needs to examine many clusters in order to get the gist of the topic. At the limit, singleton clusters containing just one relevant document have precision 100%, but the advantages of clustering are not realised.

Neither recall (R) nor precision (P) is a good measure of quality for ranking the clusters in the order in which the user is recommended to explore them. Recall would rank highly large clusters at the top of the hierarchy, while precision would rank highly singleton clusters containing one relevant document. Neither case is desirable for exploration. A combination of the two measures would make more sense. For convenience a measures

whose values range from 0 (meaning worst case, i.e. $R = 0$ and $P = 0$) and 1 (meaning best case, i.e. $R = 1$ and $P = 1$) would be good, which suggests a mean between R and P. However, at the top of the hierarchy precision tends to be much smaller than recall ($P \ll R$), while at the bottom of the hierarchy recall tends to be much smaller ($R \ll P$). Therefore, in an additive type of mean such as arithmetic mean $\frac{R+P}{2}$ or quadratic mean $\sqrt{\frac{R^2+P^2}{2}}$, the component that is much smaller than the other is practically ignored. What one expects to happen is:

- At the top of the hierarchy precision tends to be much smaller than recall, so R is dominant in the formula. If the relevant documents are spread over several large clusters, their parent has a larger recall than any of them and will be ranked higher than all of them, so **Algorithm 6** will eliminate these clusters, as well as their successors, and only keep the parent.

- At the bottom of the hierarchy, recall tends to be much smaller than precision, precision is the dominant component, so relevant singleton clusters score much better than their parents. The algorithm will keep singleton relevant clusters and remove their parents and ancestors.

These expectations were met when we tried such formulae on our test source collection and observed an undesirable dichotomic behaviour. The partition was made up either by the root cluster, for some topics, or by the set of relevant singleton clusters. The only exception was topic 2, which has relatively few relevant documents, so the contribution to recall of each document is significant. Moreover, such a small number of relevant documents (8) tends to be concentrated low in the hierarchy, so the cluster containing them has high recall and high precision and will form the relevant cluster of the partition.

The switch, over the topics, between the two opposite types of behaviour was influenced by:

- the number of relevant documents for a topic, which determines the contribution to recall of each relevant document.

- the bias towards R or P of the quality formula, when we used weighted mean, $\frac{R+b \cdot P}{1+b}$ or $\sqrt{\frac{R^2+b \cdot P^2}{1+b}}$.

Presenting to the user one large cluster with all the relevant documents (but also many non-relevant ones) or all the documents estimated to be relevant is not desirable. Therefore, we looked for a more conservative formula, which makes sure that neither the recall nor the precision of the highly ranked clusters is too low. A candidate for such a quality measure is the harmonic mean,

$$H = \frac{2}{\frac{1}{R} + \frac{1}{P}},$$

which is biased towards the smaller of the R and P components, so the best clusters are expected to be reasonable i.e. to have sufficient recall to convey the essence of the topic, but also sufficient precision so that the topic is not too diluted. A weighted form of this formula,

$$H(b) = \frac{b+1}{b \cdot \frac{1}{R} + \frac{1}{P}}, \tag{5.1}$$

with $b > 0$, allows the user of the system to choose to see cluster with higher precision or higher recall. When $b > 1$ R is given more bias, while when $b < 1$ precision weights more.

It is apparent that another form of writing this quality formula is

$$F = \frac{(b+1) \cdot P \cdot R}{b \cdot P + R}, \tag{5.2}$$

which is exactly the complement of van Rijsbergen's E measure of effectiveness,

$$E = 1 - \frac{(\beta^2+1) \cdot P \cdot R}{\beta^2 \cdot P + R},$$

with $b = \beta^2$. Using this classic formula for measuring quality in document clustering has the advantage that old experimental results can be compared to ours or, even better, can be reviewed and re-interpreted in an interactive context. However, before examining the results obtained with this formula some observations are need:

- The F measure[2] was used for evaluating the quality of cluster-based retrieval algorithms that retrieved a single cluster. In our interactive context we want not only to find the best cluster, but to obtain and possibly to evaluate a ranked partition

[2] E and F are complementary, so conceptually they represent the same measure.

of the cluster structure. Therefore, a separate, although possibly related, measure of overall retrieval quality is required. We will deal with that issue later in this chapter.

- In our opinion, the mathematical model of F and the practical interpretation of it are largely mis-understood and mis-used. This originates from some interpretations in the paper that originally introduced F [JR71] which were never challenged or clarified.

It can be easily verified that

$$\frac{\partial F}{\partial P} / \frac{\partial F}{\partial R} = \frac{R^2}{\beta^2 \cdot P^2}, \tag{5.3}$$

from which it is obvious that

$$\beta = \frac{R}{P} \iff \frac{\partial F}{\partial P} = \frac{\partial F}{\partial R}. \tag{5.4}$$

Based on equation 5.4, Jardine and van Rijsbergen defined "the relative importance β attached to recall and precision by a user as the recall/precision ratio at which he is prepared to trade a given increment in recall for an equal loss in precision". Unfortunately, the suggested practical application and the general interpretation are that:

1. If a user gives the same importance to precision and to recall, she should set $\beta = 1$ and should expect that a variation of P alters F as much as an equal variation of R or, alternatively, that F does not change if an increase in P is matched by an equal decrease in R or vice-versa.

2. If a user gives twice as much importance to recall as to precision, she should set $\beta = 2$ and should expect that an increase in R improves F twice as much as an equal increase in P.

3. If a user gives twice as much importance to precision as to recall, she should set $\beta = 0.5$ and should expect that an increase in P improves F twice as much as an equal increase in R.

Our experiments leave no doubt that such expectations are wrong. For example, for a topic with 52 relevant documents, a cluster with 2 relevant documents out of 11 has higher F (for either of the 3 values for β) than a singleton relevant cluster, i.e. doubling the recall matters more than decreasing the precision more than 5 times. In fact, it is obvious from equation 5.3 that the relative variation of F with P and respectively with R depends not only on β, but also on the values of P and R.

Realistic expectations can be derived from equation 5.1: F is biased towards the smaller of its R and P components, so that R tends to be dominant at the bottom of the hierarchic structure, and P at the top of the structure. The parameter b can alter this balance by increasing the weight of R (when $b > 1$) or P (when $b < 1$).

- The use of the measure F for evaluating output quality is not restricted to clustering and cluster-based retrieval. In the case of best-match retrieval, for example, once the ranked list of documents is cut into the 'good' part (to be shown to the user) and the 'bad' part (to be discarded), based on a size or similarity threshold, the precision and recall of the 'good' part can be calculated and F derived as a measure of retrieval quality. Moreover, the cut-off point of the ranked list can be done based not on thresholds, but on *break-even* points, where recall and precision are equal, or where the R/P ratio is set by the user. Such points can be obtained by intersecting the P–R plot with the $P = \frac{R}{\beta}$ line, where β indicates the user's interest in recall, relative to precision.

5.4.3 Experimental results

Now let us look the result of applying **Algorithm 6** on the cluster structure with F as the cluster quality measure. **Table 5.8** shows the result for the structure obtained with the group-average clustering algorithm, when β was set to 1.0.

For each of the 6 test topics, the table presents the partition of the structure, with the clusters ranked according to $F(\beta = 1.0)$. The 'garbage' cluster, formed by merging the clusters with $F = 0$, is not shown. For each cluster, the number of documents ("Docs"), the number of relevant documents ("Rel") and the measure of quality ("F") are shown[3]. From the point of view of grouping documents relevant to each topic, it is clear from

[3]For now, the reader should ignore the last column of the table; we will deal with it later.

Topic	Docs : Rel	F($\beta = 1.0$)	Aspects
1 : 408	18 : 11	0.4151	15 16 17 19 20 21 22
	73 : 18	0.3333	0 1 2 4 5 7 8 9 11 12 13 14 17 18 23
	6 : 2	0.0976	6 16
	18 : 2	0.0755	17
	1 : 1	0.0556	10
	1 : 1	0.0556	6
2 : 414	4 : 4	0.6667	1 2 3 4 5 6 7 8 9 10 11
	1 : 1	0.2222	11
	1 : 1	0.2222	0
	1 : 1	0.2222	11
	1 : 1	0.2222	10 11
3 : 428	20 : 8	0.4000	0 1 2 3 4 5 9 11 12 13 14 17 23
	2 : 2	0.1818	16
	7 : 2	0.1481	8 15
	1 : 1	0.0952	25
	1 : 1	0.0952	24
	1 : 1	0.0952	18
	1 : 1	0.0952	23
	1 : 1	0.0952	22
	1 : 1	0.0952	19 20 21 23
	1 : 1	0.0952	10
	1 : 1	0.0952	6
4 : 431	21 : 10	0.3704	3 4 5 6 10 11 18 19 20 22 23 24 27 30 31 32 34 35 36 37 38 39
	53 : 7	0.1628	2 11 21 22 25 28 32 33 36
	4 : 3	0.1622	9 15
	6 : 2	0.1026	12 16
	1 : 1	0.0588	29
	1 : 1	0.0588	26
	1 : 1	0.0588	17
	1 : 1	0.0588	14
	1 : 1	0.0588	13
	1 : 1	0.0588	8 22
	1 : 1	0.0588	26
	1 : 1	0.0588	7
	1 : 1	0.0588	26
	1 : 1	0.0588	1
	1 : 1	0.0588	0
5 : 438	97 : 22	0.2953	2 3 4 11 12 13 15 16 17 21 28 29 30 33 34 38 39 41 42 47 49 55
	19 : 5	0.1408	31 32 36 37 43 48
	11 : 3	0.0952	0 7 35
	2 : 2	0.0741	1 19
	4 : 2	0.0714	24 40
	7 : 2	0.0678	18 22
	10 : 2	0.0645	6 27
	11 : 2	0.0635	5 23
	27 : 2	0.0506	46
	1 : 1	0.0377	50 51 52 53 54
	1 : 1	0.0377	45
	1 : 1	0.0377	44
	1 : 1	0.0377	26
	1 : 1	0.0377	25
	1 : 1	0.0377	20
	1 : 1	0.0377	14
	1 : 1	0.0377	9
	1 : 1	0.0377	10
	1 : 1	0.0377	8 53
6 : 446	41 : 12	0.3429	4 10 12
	3 : 3	0.1875	5
	9 : 3	0.1579	8 9 14
	2 : 2	0.1290	8
	2 : 2	0.1290	1 15
	1 : 1	0.0667	13
	1 : 1	0.0667	11
	1 : 1	0.0667	7
	1 : 1	0.0667	6
	1 : 1	0.0667	3
	1 : 1	0.0667	2
	1 : 1	0.0667	0

Table 5.8: Group Average, $\beta = 1.0$.

Table 5.8 that the use of F as a measure of cluster quality does a better job than the other measures tried: the best clusters do present a degree of balance between recall and precision.

For comparison, Table 5.9 shows the corresponding result of partitioning the structure built with the complete-link clustering algorithm, with $F(\beta = 1.0)$ used as measure of cluster quality. As expected, due to the different nature of the clustering algorithm, the partition is spread into a larger number of smaller clusters which tend to have lower recall and higher precision than in the case of the structure generated with group-average. There are also a larger number of singleton clusters.

The influence of β on the results is also apparent. Tables 5.10 and 5.11 show that a value of 0.5 generates a higher number of smaller clusters which tend to have higher precision and lower recall, while Table 5.12 shows that a value of 2.0 generates a smaller number of larger clusters which tend to have higher recall and lower precision.

These results offer a better view of the distribution of relevant documents over the cluster structure and give an indication of the best set of clusters that a user can hope to find when exploring the clustered document collection. However, from the point of view of the usefulness of clustering for mediation, the results are inconclusive. We do not know yet if, for example, a cluster with 18 documents, of which 11 relevant, is sufficiently good to generate a good topic model and subsequently a mediated query that can produce good retrieval effectiveness. We will have to defer such conclusion until we analyse the results of mediation experiments, in the next chapter. The best clusters obtained here will constitute the basis of those experiments.

For an operational system based on mediation, or at least on the exploration of clustered collections, the system administrator should apply several clustering methods and parameter sets and test the obtained structure with users conducting typical tasks. A variety of parameters or threshold can be set for the clustering algorithm: a maximum depth of the structure, a minimum and/or maximum number of documents in bottom clusters or a minimum and/or a maximum number of children in each cluster. The task

Topic	Docs : Rel	F($\beta = 1.0$)	Aspects
1 : 408	14 : 10	0.4082	15 16 17 19 20 21 22
	6 : 5	0.2439	7 8 9 23
	19 : 5	0.1852	11 13 14 18 23
	4 : 3	0.1538	17
	4 : 2	0.1026	0 1 2
	6 : 2	0.0976	5 10
	8 : 2	0.0930	6 16
	11 : 2	0.0870	6 17 19
	1 : 1	0.0556	17
	1 : 1	0.0556	17
	1 : 1	0.0556	12
	1 : 1	0.0556	4
2 : 414	3 : 3	0.5455	1 2 3 4 5 6 7 8 9 10 11
	1 : 1	0.2222	11
	1 : 1	0.2222	11
	1 : 1	0.2222	0
	1 : 1	0.2222	11
	1 : 1	0.2222	10 11
3 : 428	20 : 7	0.3500	1 2 3 4 5 9 11 12 13 14 17 23
	2 : 2	0.1818	16
	8 : 2	0.1429	10 19 20 21 23
	14 : 2	0.1176	6 18
	1 : 1	0.0952	25
	1 : 1	0.0952	24
	1 : 1	0.0952	23
	1 : 1	0.0952	22
	1 : 1	0.0952	15
	1 : 1	0.0952	8
	1 : 1	0.0952	0
4 : 431	6 : 5	0.2564	3 4 5 6 11 18 19 20 22 32 34 35 36 37 38 39
	15 : 4	0.1667	0 8 22 30 31 32 33 36
	17 : 4	0.1600	22 24 25 29 36
	5 : 3	0.1579	9 15
	5 : 3	0.1579	2 21 28
	2 : 2	0.1143	10 23
	5 : 2	0.1053	17 27
	6 : 2	0.1026	12 16
	1 : 1	0.0588	26
	1 : 1	0.0588	14
	1 : 1	0.0588	13
	1 : 1	0.0588	11
	1 : 1	0.0588	26
	1 : 1	0.0588	7
	1 : 1	0.0588	26
	1 : 1	0.0588	1
5 : 438	42 : 14	0.2979	2 4 7 11 13 15 17 21 28 30 38 41 47 49
	17 : 5	0.1449	10 20 22 29 48
	11 : 4	0.1270	3 18 33 44
	10 : 3	0.0968	12 39 42
	2 : 2	0.0741	31 32 37
	2 : 2	0.0741	1 19
	7 : 2	0.0678	9 16
	11 : 2	0.0635	24 40
	11 : 2	0.0635	6 27
	13 : 2	0.0615	36 45
	1 : 1	0.0377	55
	1 : 1	0.0377	50 51 52 53 54
	1 : 1	0.0377	46
	1 : 1	0.0377	43
	1 : 1	0.0377	35
	1 : 1	0.0377	34
	1 : 1	0.0377	26
	1 : 1	0.0377	23
	1 : 1	0.0377	25
	1 : 1	0.0377	14
	1 : 1	0.0377	8 53
	1 : 1	0.0377	46
	1 : 1	0.0377	5
	1 : 1	0.0377	0
6 : 446	16 : 5	0.2222	12
	19 : 5	0.2083	10 12
	3 : 3	0.1875	5
	2 : 2	0.1290	8
	2 : 2	0.1290	8 9
	5 : 2	0.1176	1 15
	1 : 1	0.0667	13
	1 : 1	0.0667	14
	1 : 1	0.0667	11
	1 : 1	0.0667	10
	1 : 1	0.0667	7
	1 : 1	0.0667	6
	1 : 1	0.0667	3
	1 : 1	0.0667	4
	1 : 1	0.0667	2
	1 : 1	0.0667	0

Table 5.9: Complete Link, $\beta = 1.0$.

Topic	Docs : Rel	$F(\beta = 0.5)$	Aspects
1 : 408	14 : 10	0.5495	15 16 17 19 20 21 22
	7 : 6	0.4762	7 8 9 23
	6 : 5	0.4237	13 14 17
	5 : 3	0.2727	0 1 2 5
	6 : 2	0.1695	6 16
	1 : 1	0.1282	17
	1 : 1	0.1282	18
	1 : 1	0.1282	17 19
	1 : 1	0.1282	17
	1 : 1	0.1282	12
	1 : 1	0.1282	11
	1 : 1	0.1282	10
	1 : 1	0.1282	6
	1 : 1	0.1282	4
2 : 414	4 : 4	0.8333	1 2 3 4 5 6 7 8 9 10 11
	1 : 1	0.4167	11
	1 : 1	0.4167	0
	1 : 1	0.4167	11
	1 : 1	0.4167	10 11
3 : 428	20 : 8	0.4000	0 1 2 3 4 5 9 11 12 13 14 17 23
	2 : 2	0.3571	16
	1 : 1	0.2083	25
	1 : 1	0.2083	24
	1 : 1	0.2083	18
	1 : 1	0.2083	23
	1 : 1	0.2083	22
	1 : 1	0.2083	19 20 21 23
	1 : 1	0.2083	15
	1 : 1	0.2083	10
	1 : 1	0.2083	8
	1 : 1	0.2083	6
4 : 431	6 : 5	0.4386	3 4 5 6 11 18 19 20 22 32 34 35 36 37 38 39
	4 : 3	0.3061	9 15
	5 : 3	0.2830	2 21 28
	2 : 2	0.2439	10 23
	6 : 2	0.1754	12 16
	1 : 1	0.1351	29
	1 : 1	0.1351	32 33
	1 : 1	0.1351	30 31 32 36
	1 : 1	0.1351	27
	1 : 1	0.1351	26
	1 : 1	0.1351	25
	1 : 1	0.1351	24
	1 : 1	0.1351	17
	1 : 1	0.1351	22 36
	1 : 1	0.1351	14
	1 : 1	0.1351	13
	1 : 1	0.1351	11
	1 : 1	0.1351	8 22
	1 : 1	0.1351	26
	1 : 1	0.1351	7
	1 : 1	0.1351	26
	1 : 1	0.1351	1
	1 : 1	0.1351	0

Table 5.10: Group Average, $\beta = 0.5$ (Part 1).

Topic	Docs : Rel	F($\beta = 0.5$)	Aspects
5 : 438	36 : 10	0.2551	2 4 11 12 21 28 29 39 42 47
	6 : 3	0.1974	13 16 38
	19 : 5	0.1953	31 32 36 37 43 48
	2 : 2	0.1667	1 19
	11 : 3	0.1563	0 7 35
	4 : 2	0.1471	24 40
	17 : 3	0.1250	3 34 55
	7 : 2	0.1250	18 22
	10 : 2	0.1087	6 27
	11 : 2	0.1042	5 23
	1 : 1	0.0893	50 51 52 53 54
	1 : 1	0.0893	49
	1 : 1	0.0893	46
	1 : 1	0.0893	45
	1 : 1	0.0893	44
	1 : 1	0.0893	41
	1 : 1	0.0893	26
	1 : 1	0.0893	33
	1 : 1	0.0893	30
	1 : 1	0.0893	25
	1 : 1	0.0893	20
	1 : 1	0.0893	15
	1 : 1	0.0893	17
	1 : 1	0.0893	14
	1 : 1	0.0893	9
	1 : 1	0.0893	10
	1 : 1	0.0893	8 53
	1 : 1	0.0893	46
6 : 446	3 : 3	0.3659	5
	4 : 3	0.3333	12
	13 : 5	0.3086	12
	2 : 2	0.2703	8
	2 : 2	0.2703	8 9
	2 : 2	0.2703	1 15
	3 : 2	0.2439	4 10
	1 : 1	0.1515	13
	1 : 1	0.1515	14
	1 : 1	0.1515	12
	1 : 1	0.1515	11
	1 : 1	0.1515	10
	1 : 1	0.1515	7
	1 : 1	0.1515	6
	1 : 1	0.1515	3
	1 : 1	0.1515	2
	1 : 1	0.1515	0

Table 5.11: Group Average, $\beta = 0.5$ (Part 2).

Topic	Docs : Rel	F($\beta = 2.0$)	Aspects
1 : 408	73 : 18	0.4225	0 1 2 4 5 7 8 9 11 12 13 14 17 18 23
	40 : 13	0.3611	15 16 17 19 20 21 22
	6 : 2	0.0685	6 16
	1 : 1	0.0355	10
	1 : 1	0.0355	6
2 : 414	4 : 4	0.5556	1 2 3 4 5 6 7 8 9 10 11
	30 : 3	0.2419	0 11
	1 : 1	0.1515	10 11
3 : 428	49 : 12	0.4651	0 1 2 3 4 5 8 9 11 12 13 14 15 17 18 19 20 21 23
	2 : 2	0.1220	16
	1 : 1	0.0617	25
	1 : 1	0.0617	24
	1 : 1	0.0617	23
	1 : 1	0.0617	22
	1 : 1	0.0617	10
	1 : 1	0.0617	6
4 : 431	21 : 10	0.3268	3 4 5 6 10 11 18 19 20 22 23 24 27 30 31 32 34 35 36 37 38 39
	53 : 7	0.1892	2 11 21 22 25 28 32 33 36
	4 : 3	0.1103	9 15
	6 : 2	0.0725	12 16
	1 : 1	0.0376	29
	1 : 1	0.0376	26
	1 : 1	0.0376	17
	1 : 1	0.0376	14
	1 : 1	0.0376	13
	1 : 1	0.0376	8 22
	1 : 1	0.0376	26
	1 : 1	0.0376	7
	1 : 1	0.0376	26
	1 : 1	0.0376	1
	1 : 1	0.0376	0
5 : 438	97 : 22	0.3607	2 3 4 11 12 13 15 16 17 21 28 29 30 33 34 38 39 41 42 47 49 55
	19 : 5	0.1101	31 32 36 37 43 48
	11 : 3	0.0685	0 7 35
	98 : 3	0.0490	44 46
	2 : 2	0.0476	1 19
	4 : 2	0.0472	24 40
	7 : 2	0.0465	18 22
	10 : 2	0.0459	6 27
	11 : 2	0.0457	5 23
	57 : 2	0.0377	8 14 53
	1 : 1	0.0239	50 51 52 53 54
	1 : 1	0.0239	45
	1 : 1	0.0239	26
	1 : 1	0.0239	25
	1 : 1	0.0239	20
	1 : 1	0.0239	9
	1 : 1	0.0239	10
6 : 446	98 : 20	0.4673	1 2 3 4 8 9 10 12 13 14 15
	73 : 5	0.1323	5 8
	1 : 1	0.0427	11
	1 : 1	0.0427	7
	1 : 1	0.0427	6
	1 : 1	0.0427	0

Table 5.12: Group Average, $\beta = 2.0$.

of the individual user may vary from just getting a gist of the topic to getting as much coverage as possible. Therefore, one can imagine that several cluster structures can be associated with each document collection, and the users can be offered guidelines as to which structures are better for specific cases such as recall-oriented or precision-oriented tasks.

5.4.4 Quality of the cluster structure

Having the quality of the cluster structure expressed as a single value would present the advantage of allowing comparison of the effect of the various independent variables. It is not possible to compare tables or graphs such as those obtained in the previous section for all combinations of parameters such as clustering methods, similarity measures, or weighting schemes. However, it is possible to compare the single quality values, either separately, for each topic, or averaged over the topics.

A serious problem, however, is finding such a measure of overall quality. A possible candidate is the *average uninterpolated precision (AUP)*, which in the case of the ranked partition can be computed, for each topic, by placing the relevant documents before the non-relevant documents in each cluster and calculating the precision at every seen relevant document. The use of a widely used measure such as AUP allows us to compare the effectiveness of a search based on the procedure described above with a 'normal' ranked retrieval, based on topic descriptions. It is worth mentioning that AUP favours systems that retrieve relevant documents quickly (early in the ranking) [BYRN99], so is particularly appropriate for interactive scenarios, in which the user wants to view as few documents as possible. It is also the preferred measure in TREC [VH00].

However, a word of caution is necessary: we decided to deliberately accept a decrease in precision in order to identify topics. Therefore, a precision-oriented measure such as AUP fails to capture the success of our endeavour. Rather, the measure can be used as a safety check: it indicates an upperbound precision that our procedure can generate. If the obtained values were too low, this would be an indication that our approach has no chances of success. In reality, we obtained decent values, as shown in the **Table 5.13.**

	Clustering Method	Group Average			Complete Link		
Topic	β	0.5	1	2	0.5	1	2
1 : 408		0.8272	0.7473	0.6624	0.7877	0.7242	0.7242
2 : 414		1.0000	1.0000	0.9036	1.0000	1.0000	0.6967
3 : 428		0.7235	0.6891	0.7229	0.9238	0.5974	0.5974
4 : 431		0.8076	0.5637	0.5637	0.7200	0.5747	1.0000
5 : 438		0.4562	0.5679	0.5342	0.5269	0.5244	0.5244
6 : 446		0.7795	0.6393	0.7507	0.6792	0.5204	0.5204
Average:		0.7277	0.7439	0.8768	0.7339	0.7059	0.8662

Table 5.13: The cluster structure quality as Average Uninterpolated Precision

A statistical analysis of variance revealed no significant effect of the parameters "Clustering method" and "Beta" on the value of AUP. A β value of 0.5 tends to give higher precision, but the difference is not statistically significant. The only significant difference was recorded over the topics, with the average uninterpolated precision being significantly higher ($p < 0.1$) for topic 2 (a more focused topic, with only 8 relevant documents) than for the other topic. However, the topics are part of the test collection and do not constitute a controlled variable (they are the random factor in the statistical analysis).

Maybe the potential success of using clustering for mediation can be better indicated by the average value of F over the clusters of the partition, for each topic. These values are presented in the **Table 5.14.**

There is no statistical difference between the results produced by different clustering methods, but $\beta = 0.5$ produced a significantly higher average F than each of the other 2 values ($p < 0.5$), between which the difference is not significant. This can be seen as an indication that weighting the cluster quality measure F towards higher precision may produce a more promising ranked partitioning for mediation. Such a hypothesis would need to be confirmed in mediation simulations.

Similar results were obtained when the other combinations of weighting scheme and similarity measure were used. On visual inspection of the results there were no apparent differences, so no formal statistical analysis was performed.

	Clustering Method	Group Average			Complete Link		
Topic	β	0.5	1	2	0.5	1	2
1 : 408		0.2175	0.1721	0.1846	0.2157	0.1328	0.0976
2 : 414		0.5000	0.3111	0.3163	0.4722	0.2761	0.2286
3 : 428		0.2367	0.1356	0.1197	0.2406	0.1326	0.1022
4 : 431		0.1686	0.0963	0.0742	0.1972	0.1057	0.1877
5 : 438		0.1139	0.0685	0.0604	0.1260	0.0666	0.0485
6 : 446		0.2104	0.1177	0.1284	0.1940	0.1038	0.0738
Average:		0.2412	0.1502	0.1473	0.2410	0.1363	0.1231

Table 5.14: The cluster structure quality as average of F

5.4.5 Aspects of relevance

From the practical perspective of using clustering for mediation the results discussed above were less successful than anticipated. The main reason is the surprisingly large number of singleton relevant clusters. According to the classic form of the cluster hypothesis, the documents relevant for each topic were expected to be highly similar and therefore to be grouped together in relevant clusters.

We decided to take a closer look at this phenomenon and to take into account aspectual relevance judgements. For each topic of the test collection the human experts had identified a number of topics and had made aspectual judgements indicating, for each relevant document, the aspects to which the document was relevant. The number of aspects for each of the 6 topics is: 24, 12, 26, 40, 56 and 16 (see **Table 5.1**) and they are identified by integers numbers starting with 0.

We extended the analysis software to produce, for each cluster in the ranked partition, the identifiers of the aspects for which the documents in the cluster were relevant. The results are presented in the column "Aspects" of the tables 5.8, 5.9, 5.10, 5.11 and 5.12. The result is striking:

- The singleton clusters relevant to a topic are most often relevant to distinct aspects of that topic. This corroborates with results of the *separation test*, presented in section 5.2, showing that the similarity between documents relevant to different aspects of the same topic is significantly lower, on average, than the similarity between documents relevant to the same aspect. The obvious explanation is that documents relevant to distinct aspects of the same topic are simply not sufficiently similar to

be grouped together by the clustering algorithm.

This result is a blow to the classic cluster hypothesis, but is a confirmation of the more relaxed *aspectual cluster hypothesis* proposed in this thesis.

- The documents that are grouped together in high quality clusters (i.e. highly topical) appear to be documents that share at least one aspect of the topic. They often seem to be documents that cover more than one aspect, which makes them more likely to be similar to other topical documents.

The aspectual cluster hypothesis, formulated based on observations during informal experiments with a variety of clustered collections was therefore formally confirmed on a sub-collection of the Financial Times part of the TREC test collection, for which aspectual relevance judgements were available. It would be desirable to produce such aspectual relevance judgements for more test collections and to reproduce our tests in order for our conjecture to be accepted by the IR research community.

5.5 Conclusion

The main achievement of this chapter is experimental support for the **aspectual cluster hypothesis:**

> *Highly similar documents tend to be relevant to the same topic. Documents relevant to the same topic may be quite dissimilar if they cover distinct aspects of the topic.*

and its **consequence:**

> *Clustering algorithms tend to group together documents that cover highly focused topics, or aspects of complex topic. Documents covering distinct aspects of complex topics tend to be spread over the cluster structure.*

This result has important implications for research in document clustering. Over the years results of clustering experiments relying on the cluster hypothesis have been inconsistent. The only investigation known to date that tried to explain such inconsistencies is Sparck Jones's [SJ73], which looked at statistical differences between document collections

on which the cluster hypothesis tests were conflicting. While her experiments failed to find any correlation between collection statistics (such as number of terms per document, per collection and per test query) and classifiability, maybe investigating the semantic complexity of test topics and of documents could prove more successful.

There are even farther-reaching implications for IR in general and in particular for the design of experiments. Some experiments such as ad-hoc TREC have attempted to use relevance feedback and other query re-formulation techniques in order to build the *one* query that achieves the best retrieval performance. For complex topics, with distinct aspects, this may be the wrong approach. No single query may be able to cover all the aspects of the topic and also achieve good precision. It may be possible that a set of queries, one for each aspect, may be more appropriate to try and develop. We will investigate this issue to some extent in the next chapter.

This chapter also has implications for mediation: we achieved a better understanding of the way relevant documents are spread over the cluster structure and of the potential of clustering for exploring a document collection. A practical result of the experiments conducted is that we have an idea of the typical number of pockets of relevance that need to be found in order to assure a good coverage of a topic as well as the typical size of good clusters that the user should be looking for.

A weakness of this chapter is the fact that the experiments were only run on one test collection, for which aspectual relevance judgements were available. On the theoretical side, these experiments will need to be repeated on different test collections in order to confirm our conclusion. On the practical side, for mediation, these experiments can provide guidelines with respect to the number and size of the best pockets of relevance. If no test topics with relevance judgements are available for a certain source collection, then a domain expert can assess the size and density of typical topics in order to provide the user with exploration guidelines.

Chapter 6

Mediation Simulations

6.1 Introduction

The *ostensive model* has been proposed for mediation: the user selects exemplary documents or clusters of documents that deal with her topic of interest, based on which the system builds a model of the topic and attempts to retrieve more documents that match this topic model. While in principle the mediation model may appear simple, implementing an operational system based on mediation is not trivial due to a number of decisions that need to be made.

Rather than just enumerating these problem-issues, let us try and put them in an operational context by looking at the typical tasks performed by users, as identified by O'Day [OJ93]:

1. **Monitoring a well-known topic or set of variables over time.**

   We can assume that the user has identified a number of exemplary documents in the source collection that are relevant to a certain topic. She expects that once these documents are bookmarked, the system can derive the topic of interest to the user and periodically search the Web or other such target collection looking for new relevant documents. Some pertinent questions can be asked:

   - How many documents need to be bookmarked in order to unambiguously define a certain topic ?

- If the documents cover several distinct aspects of a topic should the system build a complex all-encompassing topic model or a set of simple models, one for each aspect ?
- How should the mediated query be built to summarize the content of the bookmarked documents ?

2. **Following an information-gathering plan specific to the task at hand.**

   This case corresponds to a user that has experience with searching and with the domain explored, so that she can employ an analytical search strategy. From this perspective, the most obvious questions are:

   - What search strategies are best at identifying pockets of relevant documents ?

     Some competing strategies are best-match retrieval of documents, cluster-based retrieval (ranking the clusters of the structure either by matching their cluster representatives with the topic description or based on their containing a high percentage of highly-ranked documents), top-down browsing, bottom-up browsing.

   - How can search strategies on the source collection be combined with mediation in order to achieve best effectiveness in searching the target collection ?

     A variety of approaches can be imagined, such as "fuse and search" and "search and fuse", discussed in chapter 3.

3. **Exploring a topic in an undirected fashion.**

   Exploratory searches may be the only option for a user unfamiliar with a domain. Exploring a very large and unstructured target collection may be very expensive in terms of time and intellectual effort, and may not reveal the extent of a topic or the relationship between its aspects. On the other hand, exploring a relatively small and semantically structured specialized source collection is expected to be much easier and to reveal the semantic structure of the topic of interest and even of the problem domain. Moreover, if the user is able to formulate a query, not necessarily a very precise one, it is expected that the result of searching a small, specialised collection, containing a relatively high percentage of relevant documents, will be much superior to the result of directly searching the target collection. While the experiments in

the previous chapter have shown that topical documents are expected to be grouped together, some questions still remain:

- Can the users *find* relevant documents and clusters of relevant documents ?
- Should the users employ mediation as soon as they find some exemplary documents or should they attempt to build a precise and complete model of the topic of interest by comprehensively exploring the source collection before extending the search to the target collection ?

Answering all these questions requires an extensive set of simulations and user experiments, on a variety of test collections, which is outwith the scope of this thesis. What we are attempting is to start exploring these issues and shed some light on them. Apart from getting some initial results and proposing some interpretations and conclusions, we intend to highlight the areas that need further exploration and even propose methods for continuing the exploration.

This chapter attempts to address the issues above by simulating mediated searches on the Financial Times test collection already described. While separate experiments are proposed and conducted for exploring particular issues, the overall chapter should succeed in offering an indication of the potential of mediated access to improve the effectiveness of retrieval and thus in confirming or disproving the **effectiveness hypothesis** (formulated in section 3.3.1).

6.2 Topic-based searching - baseline for mediation evaluation

6.2.1 Examining search results

Mediation is proposed as an approach to improving retrieval effectiveness based on improving the query submitted to the search engine rather than the search algorithm. Our approach is successful if the queries generated through mediation produce better search effectiveness on the target collection than the queries generated by the user when no mediation is employed. In other words, we intend to compare the effectiveness of mediated searches with a baseline search. This subsection justifies and computes a baseline result,

while also presenting more arguments in favour of the use of mediation.

In our informal user experiments, when the users were given the test topics and were asked to use the search functionality of WebCluster to find relevant documents in the target collection, they invariably extracted query terms from the description of the topics. Most often it was just the title of the topic that was used, but occasionally terms were also extracted from the extended description. While query reformulation was employed by a small number of users in subsequent searches, based on the initial search results, it makes sense to derive the baseline queries from the descriptions of the test topics. Their quality, evaluated in terms of search effectiveness generated, will be compared with the quality of mediated queries.

There are more than one way of automatically generating a query from the description of the test topic and there are several parameters that may influence the search results. Our first experiment, therefore, attempts to clarify these influences.

A vector-space model was adopted as base for the experiment, with a classic dot product between queries and documents used as the matching function. This allows us to separate between term weights in the collection documents, based on collection statistics and common for the baseline and the mediated search, and term weights in the query. The term weights in the baseline query are generated based on the topic description, while the ones in the mediated query are generated based on the exemplary documents marked as relevant.

Searches were conducted both on the source collection of 747 documents, 175 of which judged relevant to at least one of the 6 test topics, and on the target collection of 210,158 documents, of which 350 had been marked relevant. The two independent variables used were:

1. the form of the topic description used for deriving a query. Three different cases were considered, according to the source of the terms making up the query:

   the title of the topic ("Title"), containing the essential keywords describing the

Weighting	Topic form	Recall	AspRecall	R-Precision	AUP
RelFreq	Title	0.930333	0.939064	0.220538	0.238420
	Description	0.978938	0.970314	0.191137	0.208351
	Full	0.978938	0.970314	0.221807	0.235148
TfIdf	Title	0.930333	0.939064	0.252169	0.246859
	Description	0.978938	0.970314	0.191056	0.222180
	Full	0.978938	0.970314	0.214095	0.242646
KL	Title	0.930333	0.939064	0.215488	0.230899
	Description	0.978938	0.970314	0.166415	0.182188
	Full	0.978938	0.970314	0.197420	0.216352

Table 6.1: Effectiveness of searching the source collection.

topic.

the description of the topic ("Description"), containing the title terms, but also offering some context of the topic or of the aspects of interest.

a combination of the two ("Full"). This form of the query contains the same terms as the "Description", but the weight of the title terms is double that of the context terms.

2. the weighting scheme used for generating the term weights of the document representations in view of best-match searching the collections. The three different schemes tested were relative frequency ("RelFreq"), tf-idf in the form used by Inquery ("TfIdf") and Kullback-Liebler ("KL"). Details about these weighting schemes were given in chapter 2.

Table 6.1 shows the results of searching the source collection, while **Table 6.2** shows the results of searching the target collection, with the measures of effectiveness averaged across the 6 topics. Let us first compare the searches across the two collections. There is no significant difference in recall (R) or aspectual recall (AspR) between searches on the two collections, indicating that the distribution of 'topical terms', specific for the relevant documents, is similar in the two collections. On the contrary, there is a highly significant difference in precision, measured both as R-precision (precision measured when the cutoff of the ranked list is equal to the number of documents judged relevant for the topic) and as average uninterpolated precision (AUP), between searching the source and the target collections:

One-way ANOVA: R-Precision versus Collection

Source	DF	SS	MS	F	P
Collection	1	0.01379	0.01379	11.41	0.004 < 0.01 (!!)
Error	16	0.01933	0.00121		
Total	17	0.03312			

Individual 95% CIs For Mean
Based on Pooled StDev

Level	N	Mean	StDev	------+-------+-------+-------
Source	9	0.20779	0.02451	(------*------)
Target	9	0.15244	0.04261	(------*------)

------+-------+-------+-------

0.150 0.180 0.210

One-way ANOVA: AUP versus Collection

Source	DF	SS	MS	F	P
Collection	1	0.05814	0.05814	56.99	0.000 < 0.01 (!!)
Error	16	0.01632	0.00102		
Total	17	0.07447			

Individual 95% CIs For Mean
Based on Pooled StDev

Level	N	Mean	StDev	--+-------+-------+-------+--
Source	9	0.22478	0.02031	(---*--)
Target	9	0.11111	0.04035	(--*---)

--+-------+-------+-------+--

0.100 0.150 0.200 0.250

Note that the use of the ANOVA test is valid in the case of R-precision and AUP as these values display a normal distribution, as indicated by an Anderson-Darling normality test. The Recall and AspRecall are not normally distributed, so a non-parametric Kruskal-Wallis test was applied instead of ANOVA. In the rest of the chapter the result of the normality test will not be mentioned. For normal distribution of results the parametric ANOVA test will be used; otherwise, the non-parametric Kruskal-Wallis test will

Weighting	Topic form	Recall	AspRecall	R-Precision	AUP
RelFreq	Title	0.913373	0.948451	0.159715	0.122101
	Description	0.975763	0.979701	0.058205	0.035575
	Full	0.975763	0.979701	0.139102	0.097903
TfIdf	Title	0.913373	0.948451	0.198425	0.158780
	Description	0.975763	0.979701	0.177197	0.121106
	Full	0.975763	0.979701	0.187624	0.155078
KL	Title	0.913373	0.948451	0.166043	0.130101
	Description	0.975763	0.979701	0.119577	0.062853
	Full	0.975763	0.979701	0.166043	0.116511

Table 6.2: Effectiveness of searching the target collection.

be used instead.

The highly significant difference in precision shown above adds to the argument in favour of mediation: when searching a relatively small, specialised source collection based on a more or less vague description of an information need, a user has good chances of finding a high percentage of relevant documents after examining several tens of top-ranked documents. On a larger and more heterogeneous target collection the user may have to scan several hundreds or thousands of top-ranked documents in order to make sure that most relevant documents have been retrieved. What we expect to achieve through mediation is to improve the query used for searching the target collection so that the precision of the search increases substantially, so that user's effort of scanning the ranked list of hits from the target collection is diminished.

Now let us compare the effect of the other independent variables on the measures of effectiveness. We compared these measures both separately, on each of the two collections, and overall. The results were quite similar, so we will only report the overall measurements.

As no threshold was imposed on the ranked list of retrieved documents (which means that all documents that contained at least one query term were retrieved), R and AspR were not influenced by the weighting scheme used for searching.

There is a highly significant variation of R and AspR with the form of the topic used for building the query, in the sense that the context terms decisively contribute to recall:

Kruskal-Wallis Test on Recall

TopicForm	N	Median	Ave Rank	Z
Description	6	0.9774	12.5	1.69
Full	6	0.9774	12.5	1.69
Title	6	0.9219	3.5	-3.37
Overall	18		9.5	

H = 11.37 DF = 2 P = 0.003 < 0.01 (!!)

H = 12.36 DF = 2 P = 0.002 (adjusted for ties) < 0.01 (!!)

Kruskal-Wallis Test on AspRecall

TopicForm	N	Median	Ave Rank	Z
Description	6	0.9750	12.5	1.69
Full	6	0.9750	12.5	1.69
Title	6	0.9438	3.5	-3.37
Overall	18		9.5	

H = 11.37 DF = 2 P = 0.003 < 0.01 (!!)

H = 12.36 DF = 2 P = 0.002 (adjusted for ties) < 0.01 (!!)

The interpretation is that adding an extended description of the topic to the topic title does increase R and AspR significantly, indicating that some relevant documents match the context description, but not the title of the topic. This result confirms one of the intuitive ideas behind mediation: enriching the query with some context can significantly improve recall (both absolute and aspectual). This is important for tasks where recall is essential and also for exploratory searches, when a user unfamiliar with the domain may find it difficult to produce a query that comprehensively conveys her information need.

It is worth noting that R and AspR are quite high on both source and target searches, indicating that most relevant documents contain the keywords used for describing the topic. However, the values are less than 1, which indicates that there are relevant documents that do not contain any of the terms used for describing the topic. For example, in the case of the first test topic, the results show that there are documents about tropical

storms that have caused property damage or loss of life, but do not contain any of the words "tropical", "storm", "property", "damage", "loss" and "life". We would expect that through mediation even such high levels of recall can be improved.

A somewhat surprising result was obtained when examining the effect of the topic form on precision. We had expected that using an extended description of the topic for producing a query would generate better precision (because more context was being provided) and that using the full description (combination of title and extended description) would improve precision even more (because context terms were being used, but the title terms were being given higher weight). The analysis of variance indicates a consistent, although not quite significant, advantage of using just the title. The full description is slightly worse, while the simple description comes a more distant third. (A statistically significant difference in R-precision is obtained between "Title" and "Description" when the analysis is conducted separately only on the source collection.) An explanation for this result is apparent from more closely examining the topics and the relevant documents: the context offered by the topic description is worded so that it would help a human decide whether a certain document is relevant or not, but it does not offer, in general, terms that are expected to be found in the relevant documents. In other words, some of the descriptive terms are not content-bearing. So, the challenge for mediation is to produce a query that has a high number of terms that are specific for relevant documents, but not for non-relevant documents.

The consistent, although not quite statistically significant, advantage of the combination ("Title" + "Description") over "Description" in terms of precision also indicates that the weight of the query terms is important. More evidence in support of this statement is expected from the mediation experiments.

A consistent, although not statistically significant, effect on precision was also obtained for the weighting schemes[1]. TfIdf gave better precision (both R-precision and AUP) than the other two weighting schemes, with Kullback-Liebler and relative frequency having

[1]By applying various weighting scheme but no output cut-off, the set of retrieved documents is the same, so the recall does not change. What changes is the ordering of the hits and, consequently, the precision.

roughly equal effects. It is interesting to observe that KL was better than TfIdf at distinguishing topics in the cluster hypothesis experiment, but is not as good in terms of best-match searching. One explanation may be that KL is better at highlighting differences between documents, or between documents and queries, while TfIdf is better at highlighting similarities. It is also worth mentioning that research in using language models applied to IR is still in early stages, so the pure, theoretical form of the KL formula was employed in these experiments, while TfIdf was used in the tuned form, obtained after many years of TREC experiments with Inquery.

A somewhat disappointing aspect of these results is the generally low values of precision. This, corroborated with the rather low similarity values even between documents relevant to the same topic, as shown in the previous chapter, indicates a rather poor quality of the test topics. Having rather vague topics is worrying, as the topic models that we hope to build for mediation will consequently be rather vague. Unfortunately, no better test collection was available to simultaneously support experiments on the aspectual cluster hypothesis, simulations of mediation, and user experiments. One consequence is that results and conclusions reported here should be verified on other collections before being generalised. On the other hand, if mediation experiments are successful on such poor topics, than we are encouraged to believe that they would be much better on a better collection and, moreover, that mediation is likely to be very successful when used with a high quality specialised collection.

6.2.2 Residual effectiveness as baseline

The results above paint a good picture of the search process, comparing the effect of certain parameters on the retrieval effectiveness and also comparing the effectiveness of searching a relatively small source collection against searching a large and heterogeneous target collection. However, the results obtained are not appropriate as a baseline for evaluating the (expected) increase in performance produced by mediation. This is because in evaluating the retrieval effectiveness we considered all the documents marked as relevant in the target collection, including those that also occur in the source collection. If the mediated query, generated based on the documents marked relevant in the source, leads to the retrieval of the same documents in the target collection, the process is not very

Weighting	Topic form	Recall	AspRecall	RelAspR	R-Precision	AUP
RelFreq	Title	0.895873	0.427541	0.949383	0.085108	0.073193
	Description	0.972222	0.444902	0.993827	0.045604	0.027714
	Full	0.972222	0.444902	0.993827	0.088580	0.051533
TfIdf	Title	0.895873	0.427541	0.949383	0.128600	0.098780
	Description	0.972222	0.444902	0.993827	0.098628	0.078423
	Full	0.972222	0.444902	0.993827	0.105492	0.099499
KL	Title	0.895873	0.427541	0.949383	0.097118	0.075118
	Description	0.972222	0.444902	0.993827	0.077018	0.043545
	Full	0.972222	0.444902	0.993827	0.121424	0.063922

Table 6.3: Residual effectiveness of searching the target collection, baseline for mediation experiments.

helpful. In evaluating the effectiveness of the mediation, we should only take into account the newly found documents, which are not in the source collection. In fact, as explained in section 3.4.2, the source collection was built in view of these mediation experiments: it contains half (175) of the 350 relevant documents, the other 175 (which we will call the *target relevant documents*) being envisaged to be used for evaluating *residual* effectiveness of searches on the target collection. Another way of viewing this setting is that half of the relevant documents are used for training a topic model and the other half for testing the quality of the topic model.

It is debatable whether the non-relevant documents from the source collection should also be excluded when estimating the effectiveness of mediated search. It can be argued that, if the user rejects non-relevant documents from the source collection, these documents should be filtered out if retrieved from the target collection. However, our current model of mediation is only based on positive feedback from the user, and the scenarios implemented by our user interfaces do not allow negative feedback. Therefore, while we leave open the possibility of extending our model in the future, we assume in the current experiments that only positive relevance judgements are possible.

Table 6.3 shows the results of evaluating the search of the target collection, with the relevant documents present in the source collection excluded from the set of documents accepted as relevant. This is the result used as **baseline** for evaluating the mediated query against the original query.

It is apparent from the data that, while the residual recall is slightly lower than the

Topic	Aspects	AspCoverage	AspRecall
1 : 408	24	6	0.250000
2 : 414	12	5	0.416667
3 : 428	26	14	0.538462
4 : 431	40	15	0.375000
5 : 438	56	27	0.482143
6 : 446	16	10	0.625000
Average:			0.447879

Table 6.4: Coverage of the aspects by target relevant documents.

recall computed when all 350 relevant documents are considered, the aspectual recall is highly significantly lower (roughly half). This is because the algorithm that divided the relevant documents into the source half and the target half was biased in favour of offering aspectual coverage in the source collection, in order to realistically simulate a specialised collection. In the case of the aspects covered by only one document, that document was allocated to the source collection. **Table 6.4** compares, for each topic, the number of aspects as established by the relevance judgements with the number of aspects actually covered by the target relevant documents.

It is clear that, even if all the target relevant documents are retrieved, the aspectual recall achieved is rather low, averaging over the topics at 0.447879. One approach to interpreting the value of aspectual recall obtained in mediation experiments is to compare it to this rather low set of upperbound values. While the solution is acceptable from the point of view of the statistical analysis, it would be difficult to interpret by someone examining the result. Therefore, we adopt a different approach, by introducing the *relative aspectual recall*, which is the ratio between the number of aspects covered by a set of retrieved documents and the number of aspects covered by the target relevant documents. The best achievable value for the relative aspectual recall is 1, so the result is easier to interpret. **Table 6.3** contains both the absolute aspectual recall, for comparison with the values in tables **6.1** and **6.2**, and its relative counterpart, for comparison with values obtained through mediation.

6.3 Nearest neighbours mediation

6.3.1 Approach

As discussed in the introductory section, of this chapter, there can be different approaches to mediation in terms of the number of documents used for generating each mediated query and the number of distinct queries submitted in a search session (for retrieving documents relevant to a certain topic). This section considers one extreme case, when each of the documents judged relevant in the source collection is used to generate a distinct query which is submitted to the target collection.

Even in this case a variety of sub-approaches can be imagined. For example the mediated queries can be generated explicitly, and a size threshold can be applied or not, and the term weights (based on the frequency of the terms in the document and in the collection) can be considered or ignored. The searches based on each query generates a ranked list, and these lists need to be fused to get an overall list of retrieved results. The fusion of these lists can be done based on the individual scores, or by considering the number of relevant documents that a hit is nearest neighbour to.

It is not our intent to compare all these possible approaches, but just to explore the use of mediation and test the hypothesis that it can improve retrieval effectiveness. Comparing and understanding the influence of various approaches or independent variables is a plus, but not necessarily essential for this project. Therefore we are going to set some specifications and parameters for this case of using individual documents for mediation. No explicit mediated query is generated, but a nearest neighbour search is used instead. Like in the case of the baseline searches, we are considering the effect of the weighting scheme on effectiveness, and are also going to verify if a variation in similarity measure generates a measurable effect.

Cosine and Dice were the two similarity measures used for computing inter-document similarities in order to find nearest neighbours, but an approximation of the classic formulae was applied. When searching for the nearest neighbours of document $X = (x_1, x_2, \ldots, x_m)$, computing its similarity to each document $Y = (y_1, y_2, \ldots, y_m)$ based on

Sim	Weighting	Recall	AspRecall	RelAspR	R-Precision	AUP
Cosine	RelFreq	1.000000	0.447879	1.000000	0.163674	0.114912
	TfIdf	1.000000	0.447879	1.000000	0.200330	0.161563
	KL	1.000000	0.447879	1.000000	0.157546	0.102566
Dice	RelFreq	1.000000	0.447879	1.000000	0.129523	0.084858
	TfIdf	1.000000	0.447879	1.000000	0.142527	0.119690
	KL	1.000000	0.447879	1.000000	0.124975	0.090645

Table 6.5: Effectiveness through nearest-neighbour mediation.

one of the classic similarity measure formulae involves a normalisation based on document sizes, $|X|$ and $|Y|$. Therefore, computing the nearest neighbours of the documents marked relevant would involve the examination of all the documents in the target collection with which the relevant documents have common terms. Even for an experimental system, but especially for an operational system, this endeavour would not be practical due to the time and memory requirements. A much faster algorithm, that only examines the relevant document X and uses the inverted file of the target collection was implemented instead, in which only the terms of Y that also appear in X were considered in the formula $|Y| = sqrt\sum(y_i)^2$. In other words, a dimension reduction of the term space is applied for each document used for mediation, taking the current document as pivot.

6.3.2 Results

Table 6.5 shows the results of the nearest neighbour mediation obtained by score-fusing the lists of nearest neighbours associated with each source relevant document. For documents that appeared in multiple lists only the higher score was considered. An immediate result is that recall and relative aspectual recall are 1 for all combinations of weighting schemes and similarity measures, indicating that all the relevant documents were retrieved. Compared to the baseline search, the recall has improved highly significantly, while the relative aspectual recall is also higher, but the difference is not statistically significant[2]:

[2]In the reports of statistical results "NN" signifies the nearest-neighbour mediation, while "Search" the baseline search based on topic descriptions.

Kruskal-Wallis Test on Recall

Approach	N	Median	Ave Rank	Z
NN	6	1.0000	12.5	3.18
Search	9	0.9722	5.0	-3.18
Overall	15		8.0	

H = 10.12 DF = 1 P = 0.001 < 0.01 (!!)

H = 11.67 DF = 1 P = 0.001 (adjusted for ties) < 0.01 (!!)

The increase in precision through nearest-neighbour mediation is highly significant:

One-way ANOVA: R-Precision versus Approach

Source	DF	SS	MS	F	P
Approach	1	0.012498	0.012498	18.74	0.001 < 0.01 (!!)
Error	13	0.008672	0.000667		
Total	14	0.021170			

Individual 95% CIs For Mean Based on Pooled StDev

Level	N	Mean	StDev	----+---------+---------+---------+-
NN	6	0.15310	0.02764	(-----*------)
Search	9	0.09417	0.02463	(----*-----)
				----+---------+---------+---------+-

Pooled StDev = 0.02583 0.090 0.120 0.150 0.180

One-way ANOVA: AUP versus Approach

Source	DF	SS	MS	F	P
Approach	1	0.007098	0.007098	10.93	0.006 < 0.01 (!!)
Error	13	0.008440	0.000649		
Total	14	0.015537			

Sim	Weighting	Recall	AspRecall	RelAspR	R-Precision	AUP
Cosine	RelFreq	1.000000	0.447879	1.000000	0.178521	0.106668
	TfIdf	1.000000	0.447879	1.000000	0.184206	0.146965
	KL	1.000000	0.447879	1.000000	0.162176	0.091494
Dice	RelFreq	1.000000	0.447879	1.000000	0.121030	0.071473
	TfIdf	1.000000	0.447879	1.000000	0.114829	0.093455
	KL	1.000000	0.447879	1.000000	0.114870	0.076433

Table 6.6: Effectiveness through nearest-neighbour mediation when only one exemplary document from each aspect is used.

```
                                        Individual 95% CIs For Mean
                                        Based on Pooled StDev
Level      N      Mean      StDev  -+--------+--------+--------+----
NN         6   0.11237   0.02759                   (-------*-------)
Search     9   0.06797   0.02407    (-----*------)
                                    -+--------+--------+--------+----
Pooled StDev =  0.02548          0.050     0.075     0.100     0.125
```

We also considered the scenario when the user does not do a comprehensive search for relevant documents in the source selection, but is satisfied (possibly because of a time restriction) with finding just one exemplary document in each 'pocket of relevance'. We simulated this scenario by taking just one relevant document for each aspect of each topic and using the obtained set for nearest neighbour mediation. The results are in Table 6.6, with "AspNN" standing for aspectual nearest-neighbour mediation. The statistical analysis of variance shows that the improvement in precision, compared to the baseline, is still highly significant, although slightly lower than when all relevant documents are used (The difference of precision between the full form of nearest-neighbour mediation and the aspectual form is not significant.):

One-way ANOVA: R-Precision versus Approach

Source	DF	SS	MS	F	P
Approach	1	0.009646	0.009646	12.30	0.004 < 0.01 (!!)
Error	13	0.010194	0.000784		
Total	14	0.019841			

```
                                          Individual 95% CIs For Mean
                                          Based on Pooled StDev
Level      N      Mean     StDev  -----+--------+--------+--------+
AspNN      6   0.14594   0.03269                  (-------*-------)
Search     9   0.09417   0.02463   (----*-----)
                                   -----+--------+--------+--------+
Pooled StDev =  0.02800            0.090    0.120    0.150    0.180
```

One-way ANOVA: AUP versus Approach

```
Source      DF        SS        MS        F        P
Approach     1  0.003192  0.003192     4.98  0.044 < 0.05 (!)
Error       13  0.008339  0.000641
Total       14  0.011531

                                          Individual 95% CIs For Mean
                                          Based on Pooled StDev
Level      N      Mean     StDev  -----+--------+--------+--------+
AspNN      6   0.09775   0.02722                  (---------*---------)
Search     9   0.06797   0.02407   (-------*-------)
                                   -----+--------+--------+--------+
Pooled StDev =  0.02533            0.060    0.080    0.100    0.120
```

Recall and relative aspectual recall are 1, so even using just one exemplary document for each aspect will support the retrieval of all relevant documents from the target collection.

Let us now look at the influence of the independent factors on the precision of mediation[3]. Although there is no statistical significant difference, TfIdf consistently generates higher precision (both in terms of R-precision and AUP). This result correlates with the slight superiority of TfIdf in query-based searching. A better marked difference can be observed in the influence of the similarity measure on precision:

[3]The results reported were obtained when analysing together the effectiveness values obtained through absolute and aspectual nearest-neighbour mediation.

One-way ANOVA: R-Precision versus Sim

Source	DF	SS	MS	F	P
Sim	1	0.007435	0.007435	39.49	0.000 < 0.01 (!!)
Error	10	0.001883	0.000188		
Total	11	0.009318			

Individual 95% CIs For Mean
Based on Pooled StDev

Level	N	Mean	StDev	-----+--------+--------+--------+
Cosine	6	0.17441	0.01633	(----*---)
Dice	6	0.12463	0.01048	(---*---)
				-----+--------+--------+--------+
Pooled StDev =		0.01372		0.125 0.150 0.175 0.200

One-way ANOVA: AUP versus Sim

Source	DF	SS	MS	F	P
Sim	1	0.002933	0.002933	5.62	0.039
Error	10	0.005217	0.000522		
Total	11	0.008151			

Individual 95% CIs For Mean
Based on Pooled StDev

Level	N	Mean	StDev	-----+--------+--------+--------+
Cosine	6	0.12069	0.02747	(--------*--------)
Dice	6	0.08943	0.01700	(--------*--------)
				-----+--------+--------+--------+
Pooled StDev =		0.02284		0.080 0.100 0.120 0.140

In terms of average uninterpolated precision the superiority of Cosine is not statistically significant, but it is highly significant in terms of R-precision, indicating that Cosine is much better than Dice at identifying relevant documents at the top of the ranked list. This is essential for interactive retrieval, when the user tends to only scan the top hits.

6.3.3 Discussion

The experimental results, simulating nearest-neighbour mediation, clearly indicate a highly significant improvement in retrieval effectiveness compared to the baseline, even if the user only selects one exemplary document for each aspect of the topic investigated. In terms of usefulness and usability of this procedure, the nearest neighbour mediation is more computationally intensive than direct searching, so the requirements in computer memory and the time required for searching a large collection are increased.

It is interesting to note the effect of the search parameters (the independent variables in the experiment) on the search effectiveness. It appears that there is a negative correlation between the capacity of a parameter to separate topics in a collection (presumably based on better highlighting differences between documents) and its capacity for effective best-match retrieval (presumably based on better highlighting similarities between documents or between documents and queries). For example, KL and Dice were shown in the previous chapter to be better in clustering experiments, while here TfIdf and Cosine proved superior in searching.

6.4 Topic Models for mediation

6.4.1 Upperbound experiment

We now intend to investigate the capacity of statistical language models to support topic models, built based on a set of exemplary documents judged relevant, and to generate queries for mediation. Before exploring mediation through a real cluster structure, obtained with concrete clustering methods, we are starting with an upperbound experiment: we consider an 'ideal' clustering method that groups together in one cluster, for each topic, all the relevant documents and no non-relevant documents. While in practice this case is unlikely to happen, this experiment allows us to analyse the effect on retrieval effectiveness, and therefore on mediation quality, of various independent variables.

The reader is reminded that our representation of a topic model is a vector of weighted terms, the weights indicating the contribution of each term to the topic. Basically, the top-ranking terms of the topic are those that are very specific to the documents relevant

QuerySize	Weighting	Recall	RelAspR	R-Precision	AUP
100	RelFreq	1.000000	1.000000	0.149928	0.111442
	TfIdf	1.000000	1.000000	0.178181	0.143192
	KL	1.000000	1.000000	0.150785	0.111352
75	RelFreq	1.000000	1.000000	0.118718	0.109929
	TfIdf	1.000000	1.000000	0.172434	0.142281
	KL	1.000000	1.000000	0.150785	0.110207
50	RelFreq	1.000000	1.000000	0.139551	0.109085
	TfIdf	1.000000	1.000000	0.172434	0.140619
	KL	1.000000	1.000000	0.139290	0.108912
40	RelFreq	1.000000	1.000000	0.133804	0.108870
	TfIdf	1.000000	1.000000	0.172434	0.141274
	KL	1.000000	1.000000	0.133543	0.109140
30	RelFreq	0.994253	0.983333	0.139551	0.110003
	TfIdf	0.994253	0.983333	0.172434	0.141272
	KL	0.994253	0.983333	0.133543	0.110414
20	RelFreq	0.994253	0.983333	0.136079	0.107119
	TfIdf	0.994253	0.983333	0.166687	0.137744
	KL	0.994253	0.983333	0.139290	0.108946
15	RelFreq	0.994253	0.983333	0.124584	0.105247
	TfIdf	0.994253	0.983333	0.166687	0.136986
	KL	0.994253	0.983333	0.138445	0.108360
10	RelFreq	0.988506	0.983333	0.124584	0.104713
	TfIdf	0.988506	0.983333	0.166687	0.138048
	KL	0.988506	0.983333	0.144192	0.108114
5	RelFreq	0.969987	0.955556	0.094532	0.101207
	TfIdf	0.969987	0.955556	0.151680	0.131001
	KL	0.969987	0.955556	0.112710	0.104684

Table 6.7: Effectiveness of mediated search based on topic models.

to the topic (i.e. appear frequently in them), but are not specific to other documents. In chapter 3 we described the use of the Kullback-Liebler divergence in comparing frequency distributions of terms so that we can estimate the specificity of each term for a set of documents, relative to a corpus.

The source collection used in our experiment is relatively heterogeneous, in that it does not cover a specific domain. Therefore, we are employing a simple formula, similar to the one used to generate absolute cluster representatives (see formula 3.3 in section 3.2):

$$w_i = KL_i(RelSet, Corpus) = p_{i,RelSet} \log \frac{p_{i,RelSet}}{p_{i,Corpus}},$$

with the weight w_i of term i obtained from the frequency distribution of the term in the set of relevant documents, and in the corpus. In order to generate a mediated query from the topic model, the terms are ranked based on their weight and a cutoff is applied, usually

according to the intended query size.

The experiment reported in this section consists in taking, for each topic, the set of all relevant documents in the source collection, generating a topic model for it, and deriving the mediated query to be submitted to the target collection. The independent variables considered are the size of the query derived from the topic model and the weighting scheme used in the searching process. The results are displayed in **Table 6.7.**

The increase in precision of retrieval due to mediation, compared to the baseline search, expressed both as R-precision (which is normally distributed) and as AUP (which is not normally distributed) is highly significant, as shown by an ANOVA and respectively a Kruskal-Wallis test:

One-way Analysis of Variance for R-Precision

Source	DF	SS	MS	F	P
Approach	1	0.017655	0.017655	36.93	0.000 < 0.01 (!!)
Error	34	0.016256	0.000478		
Total	35	0.033912			

Individual 95% CIs For Mean
Based on Pooled StDev

Level	N	Mean	StDev	--------+--------+--------+------
Search	9	0.09417	0.02463	(----*----)
Topic	27	0.14532	0.02094	(--*--)
				--------+--------+--------+------
Pooled StDev =	0.02187			0.100 0.125 0.150

Kruskal-Wallis Test on AUP

Approach	N	Median	Ave Rank	Z
Search	9	0.07319	5.0	-4.44
Topic	27	0.11000	23.0	4.44
Overall	36		18.5	

H = 19.70 DF = 1 P = 0.000 < 0.01 (!!)

The increase in absolute recall is also highly significant. A mediated query of as little as 5 terms gives better recall than a query based on the topic title, while longer mediated queries consistently generate better recall than a query based on the full description of the topic. Aspectual recall, already close to 1 in the baseline experiment when the full description of the topics is used as a query, only shows a significant improvement for longer mediated queries, which cover all the relevant aspects:

Kruskal-Wallis Test on Recall

Approach	N	Median	Ave Rank	Z
Search	9	0.9722	7.0	-3.78
Topic	27	0.9943	22.3	3.78
Overall	36		18.5	

H = 14.30 DF = 1 P = 0.000 < 0.01 (!!)

H = 15.18 DF = 1 P = 0.000 (adjusted for ties) < 0.01 (!!)

Kruskal-Wallis Test on RelAspR

Approach	N	Median	Ave Rank	Z
Search	9	0.9938	15.0	-1.15
Topic	27	0.9833	19.7	1.15
Overall	36		18.5	

H = 1.32 DF = 1 P = 0.250

H = 1.44 DF = 1 P = 0.230 (adjusted for ties)

Neither the increase in precision or in recall is as pronounced as when nearest-neighbour mediation is used. However, there are some advantages of the topic-mediated approach:

- It is computationally much less demanding to build a topic model from a set of documents and to do just one search than to do a set of nearest-neighbour searches and to fuse the results. An important contributor to this difference is the fact that the time needed for searches based on inverted files increases linearly with the size of the query, so nearest-neighbour searches of long documents take much longer than searches based on relatively short queries.

	R-Precision		AUP	
	Correlation	p-value	Correlation	p-value
RelFreq	0.537	0.136	0.819	0.007
Tf-idf	0.759	0.018	0.782	0.013
KL	0.662	0.052	0.747	0.021

Table 6.8: Correlation between precision and query size in topic-based mediation.

- Expressing a topic as a ranked list of terms weighted according to their specificity to the topic may better convey the topic than a set of typical documents. It is certainly faster for a user to get the gist of a topic by reading a dozen of highly topical terms than to read or at least scan through a few relevant documents. This is particularly important for bookmarking and storing topic models.

- The cutoff of the query size can be set so that it achieves a balance between high precision and especially high recall, for long queries, and speed, for short ones.

- For heterogeneous source collections the building of a topic model can have a beneficial averaging effect, while the nearest-neighbour approach can potentially favour certain aspects of the topic that may not be the most relevant for the user.

It is obvious that the longer the query the higher the recall that can be expected, although the maximum recall of 1 is reached quite quickly. The relationship between precision and the size of the query is less clear, though. In her experiments, Harman showed that for a collection there is an optimal size of a query (generated through relevance feedback) [Har92]. Her interpretation was that shorter queries do not convey sufficient details on the information need, while longer queries tend to contain irrelevant terms. In our case, due to the use of the Kullback-Liebler divergence, the topic model is made up of terms that are more specific to the relevant documents than to the overall collection. Therefore, we expect a positive correlation between the size of the query and precision.

Table 6.8 shows the results of applying a Pearson correlation test between query size and precision, separately for each weighting scheme. As anticipated, there is a consistently high positive correlation between query size and precision, although not quite statistically significant. The values are significant in the case of using TfIdf, which gives best effectiveness. In absolute values, however, the increase in precision with the size of the mediated query is not as significant as the decrease in search speed. Therefore, an operational sys-

tem would have to strike a balance between speed, on the one hand, and effectiveness, on the other hand. For example, the operational version of WebCluster typically generates mediated queries of size 20.

Now let us look at the influence of the weighting scheme on the quality of mediation. While recall is not affected if only the weights of the terms are modified and no cut-off is applied to the results, there is a highly significant influence on precision:

One-way Analysis of Variance for R-Precision

Source	DF	SS	MS	F	P
Weighting	2	0.007843	0.003922	26.42	0.000 < 0.01 (!!)
Error	24	0.003562	0.000148		
Total	26	0.011405			

Individual 95% CIs For Mean
Based on Pooled StDev

Level	N	Mean	StDev	----+---------+---------+---------+-
KL	9	0.13806	0.01145	(---*----)
RelFreq	9	0.12904	0.01607	(----*---)
TfIdf	9	0.16885	0.00748	(----*----)
				----+---------+---------+---------+-
Pooled StDev =		0.01218		0.128 0.144 0.160 0.176

Kruskal-Wallis Test on AUP

Weighting	N	Median	Ave Rank	Z
KL	9	0.1089	10.4	-1.65
RelFreq	9	0.1089	8.6	-2.52
TfIdf	9	0.1406	23.0	4.17
Overall	27		14.0	

H = 17.61 DF = 2 P = 0.000 < 0.01 (!!)

As in the previous search experiments, TfIdf is consistently and highly significantly better than the other two measures, while Kullback-Liebler tends to be slightly better than

QuerySize	Weighting	Recall	RelAspR	R-Precision	AUP
100	RelFreq	1.000000	1.000000	0.090271	0.068080
	TfIdf	1.000000	1.000000	0.102077	0.101806
	KL	1.000000	1.000000	0.077539	0.063741
75	RelFreq	1.000000	1.000000	0.091388	0.059220
	TfIdf	1.000000	1.000000	0.113651	0.097485
	KL	1.000000	1.000000	0.096370	0.059617
50	RelFreq	1.000000	1.000000	0.076774	0.046834
	TfIdf	1.000000	1.000000	0.103859	0.092796
	KL	1.000000	1.000000	0.080246	0.055344
40	RelFreq	1.000000	1.000000	0.071027	0.050412
	TfIdf	1.000000	1.000000	0.137969	0.104523
	KL	1.000000	1.000000	0.091740	0.062597
30	RelFreq	0.994253	0.983333	0.083639	0.050592
	TfIdf	0.994253	0.983333	0.169179	0.108548
	KL	0.994253	0.983333	0.116823	0.062270
20	RelFreq	0.994253	0.983333	0.080972	0.045514
	TfIdf	0.994253	0.983333	0.142518	0.104722
	KL	0.994253	0.983333	0.102662	0.056281
15	RelFreq	0.994253	0.983333	0.077619	0.045201
	TfIdf	0.994253	0.983333	0.169762	0.118708
	KL	0.994253	0.983333	0.110038	0.062268
10	RelFreq	0.988506	0.983333	0.109714	0.061200
	TfIdf	0.988506	0.983333	0.174433	0.120020
	KL	0.988506	0.983333	0.109454	0.067064
5	RelFreq	0.969987	0.955556	0.076149	0.061324
	TfIdf	0.969987	0.955556	0.120226	0.102199
	KL	0.969987	0.955556	0.074652	0.063349

Table 6.9: Effectiveness of mediated search based on topic models when query weights are ignored.

relative frequency. Therefore, unless smoothing techniques are perfected for the relatively new KL formula, in order to improve its effectiveness, TfIdf will be the weighting scheme of choice for our operational mediation system.

6.4.2 The effect of query term weighting

In the previous experiment we assumed that the search engine used for retrieval on the target collection can deal with weighted queries. However, that assumption cannot be taken for granted: there are a large number of engines, including Web search engines, that do not allow query weights. Therefore, we are investigating here the difference in effectiveness between weighted and un-weighted mediated queries. Moreover, we intend to establish whether 'un-weighted mediation' still improves search effectiveness.

We compared the difference in performance between the two cases by repeating the

experiment presented in the previous subsection, but with all the query weights set to 1.0 (after the real weights were used to rank the terms and to choose the most significant ones). Table 6.9 shows the new result. As expected, the statistical analysis indicates a marked difference between the two cases (with "Topic" indicating topical mediation and the "NW" ending indicating no query weights):

Kruskal-Wallis Test: R-Precision versus Approach

Approach	N	Median	Ave Rank	Z
Topic	27	0.1396	37.0	4.45
TopicNW	27	0.1021	18.0	-4.45
Overall	54		27.5	

H = 19.84 DF = 1 P = 0.000 < 0.01 (!!)

H = 19.86 DF = 1 P = 0.000 (adjusted for ties) < 0.01 (!!)

Kruskal-Wallis Test: AUP versus Approach

Approach	N	Median	Ave Rank	Z
Topic	27	0.11000	39.2	5.46
TopicNW	27	0.06260	15.8	-5.46
Overall	54		27.5	

H = 29.79 DF = 1 P = 0.000 < 0.01 (!!)

Is is clear that the weights are decisive in generating a substantially improved precision (the recall is not affected if no output cutoff is applied). Therefore, search engines that allow query terms weighting should be preferred. Probably even just separating the query terms into *essential, topical terms* and *context definition terms* and assigning two distinct levels of weighting could improve mediation significantly, as shown in previous work [Har80].

Let us now compare the precision of "un-weighted mediation" with that of the baseline search. We cannot have a valid comparison with the search based on "Full" description of the topic, which was based on a weighted query, but only with the searches based on

"Title" (containing the most topical terms) and "Description" (also offering context), for which the query terms were equally weighted.

Compared to the baseline search based on the topic title, there is no statistically significant difference in precision (either RP or AUP). Compared to the baseline search based on the topic description, there is a significant increase in both forms of precision:

Kruskal-Wallis Test on R-Precision

Collection	N	Median	Ave Rank	Z
SearchD	3	0.07702	7.7	-1.62
TopicNW	27	0.10208	16.4	1.62
Overall	30		15.5	

H = 2.64 DF = 1 P = 0.104 < 0.05 (!)

Kruskal-Wallis Test on AUP

Collection	N	Median	Ave Rank	Z
SearchD	3	0.04354	8.0	-1.56
TopicNW	27	0.06260	16.3	1.56
Overall	30		15.5	

H = 2.42 DF = 1 P = 0.120 < 0.05 (!)

In conclusion, mediation is less useful when no query term weighting is allowed for searching the target collection, but is still useful: the increase in recall is highly significant, without a loss in precision.

6.4.3 The effect of term frequency uniformity

In the simple formula for generating term weights for the topic model, used in section 6.4.1, the set of relevant documents was modelled as a *bag of terms*, and the spread of each term over these documents was ignored. Here we are investigating whether the distribution of terms, not only their frequency, should be taken into account when generating

QuerySize	Weighting	Recall	RelAspR	R-Precision	AUP
100	RelFreq	1.000000	1.000000	0.160152	0.111076
	TfIdf	1.000000	1.000000	0.184921	0.153494
	KL	1.000000	1.000000	0.143103	0.104600
75	RelFreq	1.000000	1.000000	0.150348	0.111036
	TfIdf	1.000000	1.000000	0.175662	0.151738
	KL	1.000000	1.000000	0.143103	0.105010
50	RelFreq	1.000000	1.000000	0.145718	0.107599
	TfIdf	1.000000	1.000000	0.195378	0.151408
	KL	1.000000	1.000000	0.143103	0.104328
40	RelFreq	1.000000	1.000000	0.150620	0.107019
	TfIdf	1.000000	1.000000	0.195378	0.148983
	KL	1.000000	1.000000	0.132726	0.101302
30	RelFreq	0.994253	0.983333	0.125810	0.095703
	TfIdf	0.994253	0.983333	0.169955	0.143216
	KL	0.994253	0.983333	0.141060	0.099125
20	RelFreq	0.994253	0.983333	0.122338	0.089971
	TfIdf	0.994253	0.983333	0.165325	0.144179
	KL	0.994253	0.983333	0.141060	0.098688
15	RelFreq	0.988506	0.983333	0.120908	0.085399
	TfIdf	0.988506	0.983333	0.154909	0.138108
	KL	0.988506	0.983333	0.130410	0.093854
10	RelFreq	0.988506	0.983333	0.105942	0.078988
	TfIdf	0.988506	0.983333	0.144532	0.127810
	KL	0.988506	0.983333	0.120034	0.091292
5	RelFreq	0.974617	0.983333	0.087463	0.072859
	TfIdf	0.974617	0.983333	0.147743	0.118424
	KL	0.974617	0.983333	0.111973	0.087323

Table 6.10: Effectiveness of mediated search based on topic models with the uniformity/variability of the term frequencies taken into account.

the topic model.

If a term $t1$ appears once in each of the, say, k documents relevant to a certain topic, while a term $t2$ appears k times in one of the relevant documents and does not appears in the others, their frequency in the relevant set is the same. Intuitively, $t1$ can be expected to be somewhat specific to the topic as a whole, while $t2$ is expected to be highly relevant to a certain aspect of the topic, but not to the others. We want to investigate whether the variability (or, on the contrary, the uniformity) of term frequency is important or, in other words, whether it makes any difference between considering the relevant documents a set of term bags and considering it one large bag of terms.

We ran again the same mediation experiment, but modified the weights of the query terms by dividing it by $1 + \sigma$, σ being the standard deviation of the term frequency in

the documents marked relevant. The results are in **Table 6.10.** The statistical analysis showed no significant difference:

One-way ANOVA: R-Precision versus Approach

Source	DF	SS	MS	F	P
Approach	1	0.000004	0.000004	0.01	0.936
Error	52	0.028864	0.000555		
Total	53	0.028867			

Level	N	Mean	StDev	Individual 95% CIs For Mean Based on Pooled StDev
Topic	27	0.14532	0.02094	(----------*----------)
TopicU	27	0.14480	0.02591	(----------*----------)

Pooled StDev = 0.02356

0.1380 0.1440 0.1500 0.1560

Kruskal-Wallis Test: AUP versus Approach

Approach	N	Median	Ave Rank	Z
Topic	27	0.1100	31.1	1.69
TopicU	27	0.1050	23.9	-1.69
Overall	54		27.5	

H = 2.85 DF = 1 P = 0.092

Our intuition was not confirmed by the experiment. Although it may be possible that our intuition was wrong, other explanations for this result are also possible: our source collection was too small for the effect of the frequency uniformity to be measurable; or the formula we used was not sufficiently good to distinguish between documents with similar frequency, but different distribution. We attempted to modify this formula, by modifying σ's coefficient, but the results were inconsistent. Hopefully future work will help clarify this aspect.

Maybe a clearer effect of the term uniformity can be obtained in 'real-life' mediation, where there is an error factor introduced by documents wrongly marked relevant by the searcher. In such a situation taking into account term frequency uniformity may diminish the error and consequently improve mediation efficacy.

6.4.4 A conclusion for the upperbound experiment

Our experiments, although limited to one test collection and a small number of test topics, have shown the potential of mediation in improving retrieval effectiveness. Our results indicate that mediation can be used both as a *recall device*, practically bringing recall and aspectual recall to or close to 1, and as a *precision device*, significantly improving precision. Our interpretation of the result is that the topic model we obtain and the derived mediated query capture the content-bearing terms of the topic explored and is successful in ranking them according to their contribution to the topic.

We must not forget that we are dealing with an upperbound experiment, where each topic was built from all the documents marked relevant by an expert judge. It is expected that in 'real-life' mediation, if the topic is built from clusters that have a high percentage of relevant documents, the non-relevant documents will introduce an error. The error will consist of non-topical terms in the topic model, hopefully ranked at the lower end, which are expected to affect the effectiveness of mediation.

6.5 Cluster-based mediation

6.5.1 Approach

While the previous section considered the topic model of the set of all relevant documents for a certain topic, here we are investigating the potential for mediation of a real cluster structure, obtained with real algorithms. For this experiment we have at our disposal all the hierarchic structures obtained through the clustering experiments described in the previous chapter. While in the future we may want to compare the influence of clustering algorithms and their parameters for mediation, now we are content to break the ice in this kind of experiments and to consider just one structure. We chose the hierarchy produced by complete-link clustering, which is likely to identify topical clusters biased to-

QuerySize	Weighting	Recall	RelAspRecall	R-Precision	AUP
100	RelFreq	1.000000	1.000000	0.116233	0.089061
	TfIdf	1.000000	1.000000	0.112761	0.104839
	KLRel	1.000000	1.000000	0.104427	0.086998
75	RelFreq	0.982759	0.983333	0.116233	0.087721
	TfIdf	0.982759	0.983333	0.112761	0.104845
	KLRel	0.982759	0.983333	0.098680	0.087121
50	RelFreq	0.982759	0.983333	0.112761	0.088423
	TfIdf	0.982759	0.983333	0.109288	0.104558
	KLRel	0.982759	0.983333	0.107899	0.087402
40	RelFreq	0.942529	0.966667	0.112761	0.088206
	TfIdf	0.942529	0.966667	0.100955	0.105631
	KLRel	0.942529	0.966667	0.102152	0.087750
30	RelFreq	0.942529	0.966667	0.112761	0.087039
	TfIdf	0.942529	0.966667	0.112761	0.105139
	KLRel	0.942529	0.966667	0.107899	0.088020
20	RelFreq	0.898787	0.916667	0.109288	0.085861
	TfIdf	0.898787	0.916667	0.112761	0.104307
	KLRel	0.898787	0.916667	0.095208	0.087199
15	RelFreq	0.852810	0.866667	0.112761	0.088163
	TfIdf	0.852810	0.866667	0.109288	0.108837
	KLRel	0.852810	0.866667	0.104427	0.089508
10	RelFreq	0.826229	0.860494	0.097483	0.084148
	TfIdf	0.826229	0.860494	0.100955	0.103146
	KLRel	0.826229	0.860494	0.097483	0.085205
5	RelFreq	0.810105	0.816049	0.094011	0.085324
	TfIdf	0.810105	0.816049	0.097483	0.107395
	KLRel	0.810105	0.816049	0.100955	0.090245

Table 6.11: Best cluster mediation.

wards precision, Cosine as similarity measure and Kullback-Liebler as weighting scheme. In choosing the best clusters during these simulations we relied on relevance judgements and on the F measure with various values for β.

The three scenarios that we simulated are:

- The user identifies the best cluster and uses it for mediation. The simulation finds the cluster with highest F, builds a statistical model and generates a query for searching the target collection.
- The user identifies a set of top-ranking clusters that offer a coverage of all aspects and applies a "fuse and search" strategy. The simulation identifies the best set of clusters that partitions the source collection (the algorithm was described in the previous chapter), builds the overall topic model and generates the mediated query.
- The user investigates various aspects of the topic separately by identifying the set of

top-ranking clusters that cover all relevant aspects and using mediation separately, on each of them. Compared to the previous case, the simulation applies a "search and fuse" strategy: it builds a separate statistical model for each aspect, based on which it generates a query and searches the target collection, then it score-fuses the results.

For all three scenarios, the mediated queries are submitted repeatedly, truncated at various lengths in order to asses the influence the query size in retrieval effectiveness.

6.5.2 Best cluster mediation

Table 6.11 shows the result of mediating through the cluster with best F score for each topic. The variation with β of the results was negligible to at least the fifth decimal place, so only the results for $\beta = 0.5$, for precision-oriented clusters, are shown.

When compared to the baseline search, the best cluster mediation tends to generate lower recall, unless relatively long queries are employed. Even worse, aspectual recall is significantly lower:

Kruskal-Wallis Test: Recall versus Approach

Approach	N	Median	Ave Rank	Z
Clust1	27	0.9425	18.0	-0.49
Search	9	0.9722	20.0	0.49
Overall	36		18.5	

H = 0.24 DF = 1 P = 0.622

H = 0.25 DF = 1 P = 0.619 (adjusted for ties)

Kruskal-Wallis Test: RelAspR versus Approach

Approach	N	Median	Ave Rank	Z
Clust1	27	0.9667	16.3	-2.14
Search	9	0.9938	25.0	2.14
Overall	36		18.5	

H = 4.57 DF = 1 P = 0.033 < 0.05 (!)

H = 4.64 DF = 1 P = 0.031 (adjusted for ties) < 0.05 (!)

This result is hardly surprising. As clustering groups relevant document into pockets of relevance which tend to be associated with different aspects of a topic, one cannot expect even the best topic to cover all relevant aspects. On the other hand, there is an increase in precision, highly significant for the AUP measure:

Kruskal-Wallis Test: R-Precision versus Approach

Approach	N	Median	Ave Rank	Z
Clust1	27	0.10790	20.3	1.77
Search	9	0.09712	13.1	-1.77
Overall	36		18.5	

H = 3.14 DF = 1 P = 0.076

H = 3.18 DF = 1 P = 0.075 (adjusted for ties)

Kruskal-Wallis Test: AUP versus Approach

Approach	N	Median	Ave Rank	Z
Clust1	27	0.08821	21.7	3.12
Search	9	0.07319	9.0	-3.12
Overall	36		18.5	

H = 9.76 DF = 1 P = 0.002 < 0.01 (!!)

A preliminary conclusion is that, for a user employing mediation based on a clustered document collection, a one cluster strategy is fast and can be used as a precision device if the user is interested in exploring a certain aspect of an information need. If the user is interested in more than one aspect of a topic, more than one cluster should be used for mediation.

QuerySize	Weighting	Recall	RelAspRecall	R-Precision	AUP
100	RelFreq	1.000000	1.000000	0.138706	0.104265
	TfIdf	1.000000	1.000000	0.172434	0.141794
	KLRel	1.000000	1.000000	0.130332	0.104053
75	RelFreq	1.000000	1.000000	0.112970	0.104707
	TfIdf	1.000000	1.000000	0.167532	0.140122
	KLRel	1.000000	1.000000	0.136079	0.103730
50	RelFreq	1.000000	1.000000	0.133804	0.103249
	TfIdf	1.000000	1.000000	0.167532	0.139609
	KLRel	1.000000	1.000000	0.124584	0.103114
40	RelFreq	1.000000	1.000000	0.133804	0.100992
	TfIdf	1.000000	1.000000	0.171004	0.139803
	KLRel	1.000000	1.000000	0.118837	0.102751
30	RelFreq	0.994253	0.983333	0.133804	0.101347
	TfIdf	0.994253	0.983333	0.184240	0.139452
	KLRel	0.994253	0.983333	0.118837	0.103384
20	RelFreq	0.994253	0.983333	0.133804	0.097033
	TfIdf	0.994253	0.983333	0.175020	0.134235
	KLRel	0.994253	0.983333	0.118837	0.100657
15	RelFreq	0.994253	0.983333	0.118837	0.096735
	TfIdf	0.994253	0.983333	0.153724	0.133148
	KLRel	0.994253	0.983333	0.118837	0.099790
10	RelFreq	0.988506	0.983333	0.118837	0.096494
	TfIdf	0.988506	0.983333	0.160014	0.133289
	KLRel	0.988506	0.983333	0.124584	0.101134
5	RelFreq	0.969987	0.955556	0.094532	0.090278
	TfIdf	0.969987	0.955556	0.151680	0.126894
	KLRel	0.969987	0.955556	0.098004	0.097392

Table 6.12: Fuse and Search mediation.

6.5.3 Fuse and Search mediation

This strategy consists in the user selecting several clusters that are highly relevant for her information need and also offer coverage of all the aspects of interest. Our simulation builds a topic model from the fusion of the best clusters that partition the source collection, as identified by the F measure. **Table 6.12** shows the result of searching the target collection based on the query derived from the topic model built for each test topic. Again, β had no measurable effect on the results.

The strategy generates a highly significant increase in absolute recall:

Kruskal-Wallis Test: Recall versus Approach

Approach	N	Median	Ave Rank	Z
Clust	27	0.9943	22.3	3.78
Search	9	0.9722	7.0	-3.78
Overall	36		18.5	

H = 14.30 DF = 1 P = 0.000 < 0.01 (!!)

H = 15.18 DF = 1 P = 0.000 (adjusted for ties) < 0.01 (!!)

This was expected, as the partition used for mediation cover all the whole source collection, and implicitly all the aspects of each topic. However, some aspects are better represented than others in the source collection. Therefore, topical terms specific to some aspects may be ranked higher than terms specific to other aspects, in the topic model. Consequently, aspectual recall decreases through this type of mediation, unless a high number of query terms is considered:

Kruskal-Wallis Test: RelAspR versus Approach

Approach	N	Median	Ave Rank	Z
Clust	27	0.9833	19.7	1.15
Search	9	0.9938	15.0	-1.15
Overall	36		18.5	

H = 1.32 DF = 1 P = 0.250

H = 1.44 DF = 1 P = 0.230 (adjusted for ties)

The decrease is not statistically significant and is non-existent if the user set the query size high, accepting a trade-off in speed.

In terms of precision, the gain through mediation is highly significant being, for this strategy, close to that of the upperbound experiment described the previous section, when the topic was based on all the relevant documents. It appears that the error introduced in the topic models by non-relevant documents in the selected clusters does not affect precision. It is probably because, although judged non-relevant, the 'residual' documents

in each cluster are very similar, in statistical terms, to the relevant documents so the topic models are not affected greatly:

One-way ANOVA: R-Precision versus Approach

Source	DF	SS	MS	F	P
Approach	1	0.012642	0.012642	21.38	0.000 < 0.01 (!!)
Error	34	0.020102	0.000591		
Total	35	0.032745			

Level	N	Mean	StDev	Individual 95% CIs For Mean Based on Pooled StDev
Clust	27	0.13745	0.02422	(----*---)
Search	9	0.09417	0.02463	(-------*-------)

Pooled StDev = 0.02432 0.080 0.100 0.120 0.140

Kruskal-Wallis Test: AUP versus Approach

Approach	N	Median	Ave Rank	Z
Clust	27	0.10338	22.6	4.07
Search	9	0.07319	6.1	-4.07
Overall	36		18.5	

H = 16.59 DF = 1 P = 0.000 < 0.01 (!!)

If the effect of the weighting scheme on the precision of the mediated search is analysed, TfIdf significantly outperforms RelFreq and KL:

One-way ANOVA: R-Precision versus Weighting

Source	DF	SS	MS	F	P
Weighting	2	0.011853	0.005927	41.86	0.000 < 0.01 (!!)
Error	24	0.003398	0.000142		
Total	26	0.015251			

```
                                          Individual 95% CIs For Mean
                                           Based on Pooled StDev
Level      N      Mean     StDev  ----+--------+--------+--------+--
KL         9   0.12099   0.01053  (---*----)
RelFreq    9   0.12434   0.01433    (---*---)
TfIdf      9   0.16702   0.01041                (---*---)
                                  ----+--------+--------+--------+--
Pooled StDev =  0.01190          0.120    0.140    0.160    0.180
```

Kruskal-Wallis Test on AUP

Weighting	N	Median	Ave Rank	Z
KL	9	0.1028	10.6	-1.59
RelFreq	9	0.1010	8.4	-2.57
TfIdf	9	0.1395	23.0	4.17
Overall	27		14.0	

H = 17.68 DF = 2 P = 0.000 < 0.01 (!!)

6.5.4 Search and Fuse mediation

This strategy consists of the user exploring various parts of the cluster structure and selecting clusters that are relevant for different aspects of her information need. The user expects each of these clusters to be used for mediation on the target collection and the search results to be fused. Our simulation builds a topic model from each of the 'best' clusters that partition the collection (as described in the previous chapter). The query derived from each topic model is submitted to the target collection and the search results are fused.

Such an algorithm is relatively slow, especially if long queries are produced for each cluster selected for mediation. However, different clusters are expected to cover different aspects of the topic explored, so shorter queries should be sufficient. (According to the result of the best cluster mediation experiment, working with shorter queries should not degrade precision substantially.) **Table 6.13** shows the result of our simulation.

QuerySize	Weighting	Recall	RelAspR	R-Precision	AUP
20	RelFreq	1.000000	1.000000	0.024306	0.014464
	TfIdf	1.000000	1.000000	0.049961	0.020935
	KL	1.000000	1.000000	0.039312	0.016103
15	RelFreq	1.000000	1.000000	0.030053	0.010707
	TfIdf	1.000000	1.000000	0.046489	0.021164
	KL	1.000000	1.000000	0.039312	0.013578
10	RelFreq	1.000000	1.000000	0.009219	0.008634
	TfIdf	1.000000	1.000000	0.046489	0.021095
	KL	1.000000	1.000000	0.034682	0.011750
5	RelFreq	1.000000	1.000000	0.009219	0.005797
	TfIdf	1.000000	1.000000	0.036112	0.013700
	KL	1.000000	1.000000	0.010377	0.008118

Table 6.13: Search and fuse mediation.

With regards to recall, both absolute and relative recall are 1 even with queries as short as 5 terms, showing that this strategy seems to ensure a complete coverage of all aspects of interest. Precision, on the other hand, is abysmal. An explanation emerges if the log of the queries submitted to the target collection is examined: while occasional terms do suggest the original topic, most terms concentrate on a different subject. This highlights the problem of the distribution of pockets of relevance, or topic aspects, in the cluster hierarchy: if a topic does not match a major axis of the hierarchic structure, the documents relevant to it are scattered all over the hierarchy, usually in small groups. For example, documents that refer to violence against tourists are grouped around incidents that involved tourists in Egypt, Florida, Kashmir, Turkey, Mexico, Morocco, Algeria, China, and so on. Apart from having a few common terms such as "violence", and "tourist", these documents and the clusters that encompass them do not display much similarity. Therefore, generating topic models and deriving queries based on each of these clusters fails to capture the common topic. Some examples of queries generated for the "violence against tourists" topics are shown below (with the terms stemmed):

- murder charg teenag tourist brief boedek boyc stranraer glaswegian sheriff thoma world jame court connect accus new florida attempt appear
- egypt cairo islamiyya milit al claim tourist attack egyptian warn gamaa upper gama target group violenc foreign arab assiut fax
- florida collei state tourist murder miami orlando rakebrand tallahasse jagger kill impact cancell farmer rest wilhelm greg uw abta gari

- bu cairo gunmen attack islam gama assiut egyptian egypt islamiyah wound dayrut extremist tourist tour group yesterdai milit el austrian
- china taiwanes taipei taiwan mainland zhejiang talk strait boat beij suspend qiandao hangzhou chines provinc round murder hold relat march
- wound israel bethani grenad plo spanish occupi palestinian tourist end progress italian kei brief villag elect cairo talk hand west
- maasai mara bandit kenyan beaten rob briton remot tourist kenya avoid game reserv brief warn offic british attack area foreign
- eighteen hurt turkish kusadasi kurdistan explod blast guerrilla threaten tourist independ brief site bomb resort worker parti war attack part
- morocco algeria visa moroccan entri newspap border rabat insult victimis legitim simmer blown robberi lobbi shut erupt tension row disput
- istanbul attack azerbaijan lightli explos hungari guerrilla hurt kurdish wound threaten tourist immedi arm clear target campaign hit western polic
- court kidnap confess group kashmiri membership delhi alleg pakistan terrorist new suspect anti back moslem law charg brief india man

It is apparent that the commonality between these clusters is hidden by the specifics of each of them. Therefore, we attempted to capture some commonality between these disparate clusters and consequently to improve mediation precision by building an *expanded topic model*, as described in chapter 3. The intuition is that specific topics or aspects of topics tend to be part of more general topics. Therefore, the parents and ancestors of the clusters selected for mediation may capture the common topics.

We repeated the experiment described above, but built the topic model from the contribution of each selected cluster, as well as the contributions of their parents. For various decay rates of ancestor contribution we obtained various results, but the precision remained low. The conclusion is that this approach does not work if distinct relevant aspects have low similarity.

These results corroborate with the poor results obtained in the user experiment reported in section 4.6.3, indicating "search and fuse" to be a very poor mediation strategy. They also confirm, once more, our aspectual cluster hypothesis.

6.5.5 Discussion

The experiments described in this section simulated a more realistic scenario than the one underlying the upperbound experiment: rather than assuming that the user can identify all the relevant documents and use them as exemplary documents for mediation, we assumed that the user is able to identify 'good' clusters, which give a reasonable balance between recall and precision. (We will see in the next section how reasonable this assumption is.) Once such clusters are identified, the user can choose between a set of strategies, according to her specific task:

- **Best cluster** mediation - if the user is interested in quickly investigating a certain aspect of the information need.

- **Fuse and Search mediation** - if the user is interested in the overall topic and wants coverage of most relevance aspects, as well as high precision.

- **Search and Fuse mediation** - if the user is interested in exploring in more detail the particularities of various aspects of the information need.

6.6 Cluster labels for topic identification

In previous sections we assumed that the best cluster or clusters can be identified by the user in order to be used for mediation. While based on this assumption we were able to compute upperbounds of performance, we need to investigate it in order to see how well it holds and how well it can support mediation in an operational system.

For supporting the user's exploration of the clustered source collection, WebCluster uses:

relative labels, which attempt to distinguish clusters from their parent and siblings, in order to support browsing; and

absolute labels, which attempt to distinguish clusters from the rest of the collection, in order to support cluster-based searching.

Browsing can start from the top of the structured collection, or from clusters identified through cluster-based searching, or from documents identified through ranked retrieval.

It is user experiments that ultimately can establish whether the cluster labels are indicative of contents and good discriminators, for a certain collection and in a certain context. However, user tests are expensive and difficult to set up, and in general inappropriate for laboratory-type experiments that intend to check the effect on the retrieval results of various parameters. It is not feasible to change a parameter and re-do the user experiment in order to see the influence of the parameter, neither is it feasible to use factorial design in order to test the influence of all the factors. Therefore, we try to use simulations in order to estimate how likely it is that a user can identify good clusters based on their labels.

6.6.1 Absolute labels for searching

Absolute labels, built based on the Kullback-Liebler formula, attempt to convey the cluster content and at the same time to discriminate the cluster from the rest of the source document collection. When doing a cluster-based search based on the query that attempts to describe the topic of interest, the user expects the system to assign scores to clusters and to rank them so that 'good' clusters are at the top. We have argued in the previous chapter that good clusters, both for conveying information relevant to the topic of interest, and for mediation, are those with a high F score.

In an operational system there are no relevance judgements for establishing F scores for clusters. Cluster scores are usually established by matching the query (or topic description) against the cluster representatives[4]. This situation suggests the question: "Are the users able to identify good clusters through cluster-based retrieval ?".

We attempted to answer this question for our clustered source collection by checking

[4]Even if the matching process is implemented based on an inverted file structure, the formulae used are the same, and the scores obtained should be the same.

		Cosine		Dice	
Topics		NoTr	Tr20	NoTr	Tr20
1 : 408	Title	0.248995	0.215953	0.191323	0.166641
	Description	0.590561	0.559328	0.378577	0.353151
	Full	0.504004	0.464861	0.310728	0.283854
2 : 414	Title	0.330092	0.328164	0.264923	0.256465
	Description	0.314394	0.311579	0.250472	0.239510
	Full	0.331781	0.330513	0.260816	0.251055
3 : 428	Title	0.434861	0.420612	0.234689	0.218562
	Description	0.401749	0.378796	0.197111	0.185012
	Full	0.428433	0.409984	0.200752	0.187193
4 : 431	Title	0.517504	0.465249	0.519754	0.468900
	Description	0.508983	0.447143	0.433405	0.386294
	Full	0.521849	0.463817	0.474040	0.426785
5 : 438	Title	0.435968	0.395546	0.220875	0.199924
	Description	0.420173	0.383541	0.178554	0.172529
	Full	0.433475	0.394207	0.183611	0.173004
6 : 446	Title	0.340915	0.300017	0.316441	0.292495
	Description	0.291750	0.238826	0.206541	0.173350
	Full	0.330143	0.278374	0.265353	0.235637

Table 6.14: Pearson correlation between F scores and search scores.

the correlation between two sets of scores:

1. F scores, based on the relevance judgements associated with the test topics, and

2. search scores, assigned to clusters by matching their search representatives against the topic descriptions.

Therefore we computed F scores, with various values for β: 0.5 (biased towards precision), 1.0 (balanced) and 2.0 (biased towards recall) for each of the 1413 clusters in the hierarchic structure used in mediation experiments[5]. Separately, we used the Cosine and Dice similarity measures to match each the search label of each cluster against queries derived from the topic descriptions available in various forms ("Title", "Description", "All") for each of the test topics. We must stress the realism of this approach: in user experiments the users consistently built their queries based on the terms of the topic descriptions.

Table 6.14 summarizes the results of computing the Pearson correlation between the two sets of scores for each cluster, separately for each topic. The parameter β did not introduce any variation in the first 5 decimals of the results, so only the results for $\beta = 0.5$ are shown. The extra parameter considered in the experiment is the level of

[5]We tested the structure generated by the complete-link algorithm using Cosine as inter-document similarity measure and Kullback-Liebler for weighting document terms.

precision given by the size of the cluster label. In an experimental system, where speed and space requirements are not essential, we can afford storing full cluster representatives, containing all the terms that are estimated to be more specific to the cluster than to the collection as a whole. An operational system is more likely to truncate the label and only store the most highly ranked terms. In this experiment we considered both the case when no truncation was applied, and the case of the operational version of WebCluster, which stores the best 20 terms of the absolute cluster representative.

The absolute correlation values are satisfying: they indicate that cluster-based searching does indeed rank more highly the better clusters, so the **searching label hypothesis**, proposed in section 3.3.3, is confirmed for our test collection. Therefore, one of the main conditions necessary for the success of mediated access is satisfied: clustering does indeed group together topical documents and these good clusters can be identified by their most topical terms. In consequence, the user is likely to find good clusters if she is able to provide a reasonable formulation of her information need.

There is no significant difference between the various forms of the topic description. This suggests that the user does not need to make a mental effort to supply context terms for the topic; the most specific terms for the topic are sufficient to identify good clusters.

A Kruskal-Wallis statistical test (used due to the non-uniform distribution of the correlation values) shows a highly significant influence of the matching formula: Cosine does much better than Dice at identifying good clusters, so it should be the formula of choice in our operational system:

Kruskal-Wallis Test: Correlation versus Similarity

Similarity	N	Median	Ave Rank	Z
Cosine	36	0.3986	48.1	4.72
Dice	36	0.2450	24.9	-4.72
Overall	72		36.5	

H = 22.27 DF = 1 P = 0.000 < 0.01 (!!)

		Cosine		Dice	
Topics		NoTr	Tr20	NoTr	Tr20
1 : 408	Title	0.225500	0.187899	0.194860	0.164344
	Description	0.543064	0.510292	0.404397	0.372720
	Full	0.462106	0.420349	0.333977	0.300750
2 : 414	Title	0.237132	0.235300	0.221313	0.207881
	Description	0.227390	0.224768	0.212297	0.194121
	Full	0.239091	0.237645	0.222515	0.205718
3 : 428	Title	0.372372	0.354758	0.236533	0.212362
	Description	0.335100	0.313463	0.198233	0.178766
	Full	0.361774	0.342359	0.203443	0.182128
4 : 431	Title	0.431090	0.379116	0.481862	0.418612
	Description	0.421075	0.362798	0.410642	0.348078
	Full	0.433099	0.377110	0.449467	0.386120
5 : 438	Title	0.325732	0.278482	0.222716	0.189229
	Description	0.317659	0.277792	0.188103	0.169677
	Full	0.325844	0.281581	0.192057	0.169046
6 : 446	Title	0.299551	0.271031	0.318905	0.296756
	Description	0.255087	0.215404	0.220089	0.185779
	Full	0.289352	0.251286	0.281581	0.251119

Table 6.15: Pearson correlation between precision scores and search scores.

Truncating the cluster label produces results that are consistently inferior to those obtained with the full cluster label. However, the difference is not quite statistically significant, so we can conclude that the loss in cluster-based search accuracy is compensated by the gain in speed and space (memory and disk) used.

An astute reader may question the selection of F as a measure of cluster quality. It may be the case that cluster-based searching is better at identifying high-precision or high-recall clusters. We therefore repeated the experiment described above, but calculated the correlation between the precision, respectively recall of clusters, estimated based on relevance judgements, and the cluster score obtained by matching the cluster label against the topic description. **Table 6.15** and **Table 6.16** summarize the results.

As in the previous experiment, in which the F measure was considered, Cosine does highly significantly better than Dice at ranking high both high-precision and high-recall clusters. Truncating the cluster labels reduces performance, but not significantly. The most important result, however, is obtained when comparing the correlation values in the three tables. A statistical analysis of variation can be done if the correlation is viewed as the output value and the type of cluster quality (F, P, R) is viewed as one of the

		Cosine		Dice	
Topics		NoTr	Tr20	NoTr	Tr20
1 : 408	Title	0.123206	0.111896	0.080841	0.073235
	Description	0.276539	0.265344	0.145170	0.138322
	Full	0.238603	0.224422	0.118953	0.111301
2 : 414	Title	0.316941	0.313723	0.210009	0.208858
	Description	0.298420	0.293526	0.195469	0.194253
	Full	0.316889	0.313862	0.202065	0.201702
3 : 428	Title	0.237108	0.233008	0.098483	0.097368
	Description	0.225902	0.214301	0.081422	0.082120
	Full	0.237514	0.229662	0.082503	0.082558
4 : 431	Title	0.248450	0.228234	0.221692	0.208319
	Description	0.245115	0.219311	0.177926	0.168290
	Full	0.250954	0.227510	0.195788	0.186088
5 : 438	Title	0.207923	0.202449	0.065727	0.068791
	Description	0.194446	0.190875	0.045929	0.054986
	Full	0.203586	0.198936	0.049383	0.056235
6 : 446	Title	0.153286	0.130038	0.115889	0.106194
	Description	0.124906	0.100397	0.063851	0.054558
	Full	0.144798	0.118920	0.086353	0.077420

Table 6.16: Pearson correlation between recall scores and search scores.

independent variables:

Kruskal-Wallis Test: Correlation versus Quality

Quality	N	Median	Ave Rank	Z
F	72	0.3223	79.8	2.09
P	72	0.2781	65.2	-2.09
Overall	144		72.5	

H = 4.37 DF = 1 P = 0.037 < 0.05 (!)

H = 4.37 DF = 1 P = 0.037 (adjusted for ties) < 0.05 (!)

Kruskal-Wallis Test: Correlation versus Quality

Quality	N	Median	Ave Rank	Z
F	72	0.3223	99.1	7.65
R	72	0.1926	45.9	-7.65
Overall	144		72.5	

H = 58.54 DF = 1 P = 0.000 < 0.01 (!!)

It is clear that cluster-based searching, which ranks clusters according to the similarity of their absolute representative to the topic description, is significantly better at ranking highly clusters with good F than clusters with good P, and highly significantly better than ranking highly clusters with good R. The consequence for mediation is immediate: by using cluster-based searching, based on a reasonable query, the user is able to identify clusters that are good for mediation. Once these clusters are identified, the user can follow whatever mediation strategy she chooses.

Following such positive results from these correlation experiments, we also explored correlation at the high-quality end of the spectrum, by repeating the experiments and taking into account, for each topic, only clusters known to contain at least one relevant document. The attempt was to verify whether cluster-based searching is good at separating between good and very good clusters. Unfortunately, the results were inconsistent, with high correlation for some topics and no correlation or even negative correlation for other topics. The explanation lies with the quality of the topics: most of them are rather vague and poorly represented by documents in the test collection. Moreover, many documents judged relevant are only marginally relevant (but degrees of relevance are not included in the relevance judgements), which means that relevance does not necessarily correlate with a high incidence of highly topical terms in the documents, or with a good matching with the topic description.

6.6.2 Browsing labels

If the user is unfamiliar with the problem domain and is unable to formulate a query, then exploration of the hierarchic structure of source documents is possible. An appropriate visualization metaphor, coupled with informative cluster and document labels is necessary in this case.

Unfortunately, no descriptions of the various aspects for the test topics were available, so no simulation similar to the one done for searching labels is possible. Instead, we evaluated the quality of the browsing labels, and implicitly the **browsing label hypothesis**, during the user experiments described in chapter 4, and also in informal observations of users exploring various topics, during the development of our mediation prototype.

The users were unhappy with the labels displayed in the first version of WebCluster: some were very long, some were very short or empty, and many were poor indicators of content. These initial labels were based on various simple thresholding formulae: they contained terms that appeared in at least a certain number or percent of documents in the cluster. The use of language models for building the browsing labels, and more precisely the Kullback-Liebler divergence between term frequency distributions in clusters, compared to their parents, brought about a significant change in the user satisfaction. The users judged the new labels as being informative with regards to the content of the cluster and provided good support in navigation decisions. As expected, they seem to highlight which aspect of a topic is more specific in each cluster, compared to its parent cluster and to its siblings.

6.6.3 Discussion

It has been proposed that using some form of topical classification could reduce the number of documents that a user would need to view in order to acquire the needed information, and/or could contribute to the user's understanding of the semantic structure of the domain explored. Various studies attempted to verify this conjecture by comparing clustered search output with ranked retrieval [CKPW92, HPP+96, ZE98]. Most of them used simulations which assumed that the user would be able to estimate the content of the clusters based on their representative, to ignore the bad clusters, and only to explore the good ones. Few studies attempted to test this assumption, and the results of such experiments are contradictory. For example, in experiments with Scatter-Gather, Hearst and Pedersen showed that most users selected (for further exploration) the best of 5 clusters of a partition [HP96]. On the other hand, following her user studies on the use of clustered search results, Kural concluded that the users were not able to recognize good clusters based on typical document titles and topical keywords [KRJ01]. The users in her study were, however, quite successful in guessing the worst clusters, which suggests that clustering could be a good tool for space dimension reduction.

Our use of document clustering is different, but the assumption that users can find good clusters is also central to the concept of mediation. While not directly comparable

to Hearst and Kural's experiments, due to different test collections, different clustering algorithms, and different formulae for generating cluster representatives, our experiments bring a contribution to research in this area. Both simulations and user experiments have shown that, in our setting, users were able to find good clusters by combining searching and browsing. It is difficult, however, to draw definitive conclusions. More experiments are planned for the future, with different test collection and different clustering methods, for comparing the effect of a variety of formulae for generating cluster representatives: van Rijsbergen's [JR71] or Voorhees's [Voo85b] formulae used in early clustering experiments, the Robertson - Sparck-Jones or the *term selection value (TSV)* used by Kural [Kur99], the *term discrimination value* used by Dubin [Dub96], and the Kullback-Liebler-based formulae used by ourselves.

6.7 Conclusions

6.7.1 The experimental results

While we provide experimental evidence in favour of our **effectiveness hypothesis:**

> *System-based mediated access through a clustered specialised collection can improve effectiveness over un-mediated searching on a target collection.*

we are fully aware of some limitations of our experiments. The source collection, artificially built for the experiments, based on the test topics, has a particular homogeneity and level of inter-document similarity that may not appear in 'natural' collections. However, if source collections of exemplary documents were built for specialised domains, then probably their characteristics would be similar to those displayed by our collection.

More seriously, we only had 6 test topics for which relevance judgements (including aspectual relevance) were available. The results presented in this chapter were obtained by averaging effectiveness measures over the topics. Unfortunately, the variation over the topics of some of these measures was significant. Therefore we cannot generalise our result. Rather, the design, the software, the ideas generated and the conclusions of our experiments can be used as a starting point in larger scale experiments, with a larger number of source collections, of different sizes and levels of heterogeneity, and a much larger

number of test topics. However, as the quality for mediation of the six topics that were available to us was rather poor (i.e. high vagueness, low cohesion) we expect mediation on real specialised collections to be more useful than the results here would suggest. Our optimism is explained below.

It is interesting to place and compare ranked retrieval, clustering and mediation in the context of the *statistical approach* to Information Retrieval. Rather than using syntactical or semantic analysis of text, approaches specific to Natural Language Processing and, more generally, to Artificial Intelligence, we assume that the statistics of documents, and in particular the frequency distribution of terms, convey the content of the documents. Although not entirely valid, this assumption underlies all the classic retrieval models.

In this context, the best-match retrieval, the cluster hypothesis and our "mediation hypothesis" (that mediation is expected to improve retrieval effectiveness) are quite similar. In ranked retrieval, one expects documents that have the same frequency distribution of terms as the description of a topic to be relevant to that topic. In clustering, one expects similar documents (i.e. having similar statistics) to be relevant to the same topics. The same logic applies to the mediation hypothesis, which describes the expectation that documents relevant to a certain topic can make up a statistical model of the topic, which can be used to improve retrieval. If the statistical approach to IR is accepted, and the effectiveness of query-expansion methods such as relevance feedback is recognised, then the cluster hypothesis and the mediation hypothesis are logical consequences. Therefore, even if the results of our limited experiments have limited credibility, mainly due to the small sample of topics used, we are confident that these results can be confirmed and improved in the future.

6.7.2 Mediation strategies

In this chapter we simulated and compared a variety of mediation strategies: "best cluster", "fuse and search", "search and fuse". However, these were upperbound experiments, in which the best clusters, used for mediation, were chosen based on expert relevance judgements, available for each of a set of test topics. In an operational system the user is expected to combine ranked-searching, cluster-based searching and browsing in order

to identify good clusters for mediation. Once one or a set of good clusters is identified and bookmarked by the user in the source collection, a wide variety of ways to apply the mediation concept exists, apart from "fuse and search" or "search and fuse".

One idea is to follow the *information foraging* model: the user employs ranked retrieval or cluster-based searching to identify pockets of relevance corresponding to various aspects of a topic. Taking each of these 'information patches' in turn, the user browses (or forages) the neighbourhood in order to identify the cluster or set of clusters that, in her opinion, best represent that aspect of the topic investigated, and uses it to generate a mediated query and to search the target collection. Once sufficient information was gathered on the current aspect, the user can move to exploring the next aspect of relevance.

Independent of the strategy used for selecting the exemplary clusters for mediation, the user can select the mode for the mediation. In the *transparent* mode the users is shown the mediated query and can adjust it, before submitting it to the target collection, while in the *opaque* mode the system generates the query and searches the target collection in the background, on the user's request for "more documents like this". In the transparent mode, if the user has unlimited freedom to modify the query, it is difficult to use weighting of the query terms. As our mediation experiments have clearly indicated, query weighting is essential in getting good query performance. Therefore, it seems like the opaque mode is more promising from the point of view of assuring high quality mediation results. However, even in the explicit mode the user's actions could be restricted to just rejecting terms proposed by the system, so that the remaining terms will still have the weights assigned by the system, based on their frequency distribution in the documents or clusters selected as relevant. Alternatively, the more intelligent user interfaces of the future may be able to allow the user to specify the contribution of each term to the topic description.

Chapter 7

Summary and Conclusions

In this final chapter we summarize our contributions to research in Information Retrieval, discuss some limitations of our work and present our intentions for future work.

7.1 Contributions

7.1.1 System-based mediated information access - a novel interaction model for information retrieval

One main contribution of our work is proposing the concept of *system-based mediated access* as a way of emulating the human mediator, by offering the user

1. support for clarifying and refining her information need.

2. support for generating high-quality queries.

An information retrieval system based on mediation is expected to be particularly useful for novice searchers or for users exploring a new domain, with which they are unfamiliar. Such a solution is badly needed today, with Web searching tools widely available to people not trained in how to search. An automatic mediator could significantly improve searching effectiveness and, implicitly, the users' satisfaction.

Our proposed interaction model contributes to the modern interactive trend in which the user interacts with the system in order to explore a problem domain and to obtain information relevant to a certain task or information need. While related to and inspired

by other research work, our model has sufficient specificities to be considered a new interaction model.

The concept of *relevance feedback* (RF) strongly inspired our work. However, RF relies on users submitting an initial query and judging the relevance of documents in the retrieved set. We do not have this constraint: we offer a combination of browsing and searching for the user to identify documents and clusters of interest.

We were also inspired by Campbell's *ostensive model*, in which the user indicates a topic of interest by pointing to exemplary items. However, his is a query-less system which does not support the search for known items. Moreover, it provides no topical structure to support the user's exploration: the interaction relies on the user making use of similarities between documents or serendipitously finding relevant documents while browsing the collection.

WebCluster offers a combination of browsing and searching of a specialised collection in support of exploration: browsing reveals the structure of the domain, suggests search terms, and has potential for serendipitously discovering relevant documents. On the other hand, searching can reveal starting points for browsing, and even find good documents, if the searcher can formulate an adequate query. Moreover, the ranked view and the structured view of the collection are synchronised, so that a user can see the distribution of highly-ranked documents in the structure and find 'pockets' of relevant documents.

The relatively new idea of combining the hierarchic, structural view of a document collection, with a linear view, in which documents are ranked relative to a topic, was also used by Leuski. However, his system's visual interface is based on a completely different metaphor and his system is designed for exploring search results, while ours is for mediation through a (static or dynamic) specialised collection.

7.1.2 Software design

In the WebCluster project we built not only a mediation system based on clustering, but also a toolkit of components and a framework for building IR applications. This was a

major undertaking, which we hope will pay off in the long term: although in this thesis we present and analyse just one possible implementation of a mediated access system, now that the software framework is complete new applications can be built easily.

The power and flexibility of this design approach was proven by collaborative work with our sponsor, Ubilab. Our Clustering Framework was integrated by Ubilab researchers in their meta-search engine, Informia [BBMS98], in order to organise its search results. For our part, we integrated the Informia server as a search engine in user experiments that used the Web as target collection for mediation.

7.1.3 Experimental framework

The mediated access concept, proposed as a generic interaction model expected to increase retrieval effectiveness and user satisfaction, raises a multitude of theoretical and practical issues that refer to:

1. the conceptual and the mathematical model employed for representing topics.
2. the document and cluster representative formulae and the search strategies that are best for identifying relevant documents and clusters in the source collection.
3. the number of exemplary documents required to convey a topic unambiguously and the acceptable error margin.
4. the mediation strategies that offer best retrieval effectiveness.

While not attempting to solve all these issues, the thesis discussed them and proposes an evaluation methodology for investigating them. An experimental framework has been set up which consists of

1. a set of hypotheses and conjectures.
2. a set of experiment descriptions and an evaluation flow.
3. software to implement these experiments.

This experimental framework, together with the software framework that offers indexing, clustering and searching, provides the means to extend our experiments and to

easily repeat them on other test collections. Future experiments will hopefully confirm our expectations with regards to the potential of mediation to improve the effectiveness of retrieval.

7.1.4 The aspectual cluster hypothesis

The conceptual model of mediation does not impose restrictions on how the source collection should be structured. However, we investigated the use of clustering as a structuring tool and our simulations of various mediation strategies have proved the feasibility of this approach.

We also proposed the **aspectual cluster hypothesis:**

> *Highly similar documents tend to be relevant to the same topic. Documents relevant to the same topic may be quite dissimilar if they cover distinct aspects of the topic.*

and its **consequence:**

> *Clustering algorithms tend to group together documents that cover highly focused topics, or aspects of complex topic. Documents covering distinct aspects of complex topics tend to be spread over the cluster structure.*

Although only tested (successfully) on one collection, this hypothesis was actually suggested by observations on other collections. More experimentation is needed, on a variety of document collections, for this hypothesis to gain acceptance, but we are confident that the results should be positive. This result has important implications for research in document clustering as it explains inconsistencies in results of experiments relying on van Rijsbergen's original cluster hypothesis.

The fact that documents relevant to the same topics may be quite dissimilar should have a variety of implications in, for example:

- research in topic modelling, where it is assumed that topical documents have similar statistical distribution of terms.

- research in relevance feedback. Aspects of relevance should be taken into account.
- design of IR experiments. Attempting to build the one query that produces best retrieval effectiveness may not always be the right approach. For complex topics a set of queries dealing with individual aspects may perform better.

7.1.5 The use of multiple cluster representatives

Various researchers have used different formulae for generating a "one for all purposes" cluster label for representing a cluster. We have proposed a novel approach based on the fact that the label needs to distinguish a cluster in a certain context. For example, when the user is browsing, the cluster needs to be distinguished from its parent and siblings, while when the user employs cluster-based searching the best cluster needs to be distinguished from the rest of the collection. We have therefore pioneered the use of multiple representatives, according to the context (such as relative labels for browsing and absolute labels for browsing).

We have used statistical language models, and more precisely the Kullback-Liebler divergence, or relative entropy, for producing labels that distinguish the clusters from their context, by highlighting terms that are highly specific when compared to the context. The set of hypotheses proposed in section 3.3.3, namely the **browsing labels hypothesis,**

> *Browsing labels convey content and can successfully guide navigation of the source collection.*

the **searching labels hypothesis,**

> *Searching labels convey content and can successfully support search strategies in the source collection.*

and the **mediation labels hypothesis,**

> *Mediation labels can support effective search of the target collection.*

were confirmed in informal user experiments and in search simulations. The experimental results confirmed the expected capacity of the Kullback-Liebler formula to balance

accuracy with power of discrimination and to generate good document and cluster representatives.

Future experiments should be run on other test collections in order to confirm our results and hopefully to strengthen our confidence in the power of language models to convey content. Moreover, the scope of these experiments should be extended to encompass comparisons between the formulae based on language models, improved with various smoothing techniques, and traditional, heuristic formulae developed for query expansion based on relevance feedback.

7.2 Limitations

Our limited user experiments have confirmed our **usability hypothesis:**

System-based mediated access is a usable information retrieval paradigm.

and our mediation simulations have provided experimental evidence in support of our **effectiveness hypothesis:**

System-based mediated access through a clustered specialised collection can improve effectiveness over un-mediated searching on a target collection.

Although we provided an evaluation framework and conducted a relatively successful round of experiments, we have insufficient evidence to declare the success of the mediated approach to retrieval. The main weakness of our experiments is the fact that they were only run on one test collection, for which aspectual relevance judgements were available. These experiments will need to be repeated on different test collections in order to confirm our conclusions.

We also had only 6 test topics for which relevance judgements (including aspectual relevance) were available. Therefore we cannot safely generalise our results. However, these results, the conclusions drawn from them and the ideas generated can be used as a starting point in larger scale experiments, with a larger number of source collections, of different sizes and levels of heterogeneity, and a much larger number of test topics.

7.3 Future work

Firstly, the experiments described in the thesis will have to be run on more test collections. Only consistent results over a variety of collections and with different parameters can substantiate our conjectures. Moreover, user experiments are necessary, especially in an operational setting, in order to confirm the viability of our ideas and our approach.

Secondly, now that the basis for research in system-based mediated access has been set, research in a multitude of areas can contribute to better understanding and improving mediation. The following subsections discuss some directions in which we intend to continue the work started in this thesis.

7.3.1 User interfaces and visualization tools

Better user interface and visualization tools are needed for exploring the domain of interest, i.e. the source collection. We are particularly interested in combining hierarchic clustering with string-embedded algorithms: these algorithms usually consider individual documents and spread them in a low-dimension (2D) space based on the reciprocal similarities between them. However, identifying topics and navigating the obtained information space may not be easy.

An alternative would be to apply clustering to the document collection in order to obtain its topical structure. Subsequently, the spring-embedded algorithm can be applied to the cluster representatives (centroids) at a certain level of similarity in order to place them in the visualization space. The documents or subclusters can then be distributed based on their similarities to the centroids.

7.3.2 Structuring the source collection

In our experiments we have compared two clustering algorithms, complete link and group average, in terms of their capacity to identify topics and to support mediation. A possible direction of future research is to find better methods for structuring the source collection and identify topics. Possible alternatives are:

- combine existing cluster algorithms; for example complete link can be used at the

bottom of the hierarchic structure, for identifying highly specific topics, and group average or single link higher in the hierarchy, for identifying referential links between topics.

- combine document clustering techniques with language analysis to increase the probability that the fusion of similar documents identifies concepts and topics. Such an approach has been taken by Vivisimo[1].

- use the ontology / taxonomy of the specialised domain, if available, to guide the clustering, so that the emerging topical hierarchic structure is guided by the conceptual taxonomy of the domain. The advantage would be a better correlation between the automatically built structure and that expected by domain experts, and consequently a better concept learning process.

7.3.3 Document and cluster representation

We used language models in order to generate document and cluster representatives and to rank documents and clusters based on their similarity to queries. However, we employed traditional methods (Cosine and Dice) for calculating inter-document and inter-cluster similarity in the clustering process. More consistency, potentially better results, and a contribution to research in language models could result from investigating similarity models based on statistical language models.

Alternative ways of estimating the similarity between documents such as lexical affinity or word co-occurrence (bigrams) should also be investigated as they may better indicate topical similarity (and consequently produce better cluster structures) than schemes based solely on word frequency.

Future simulations of mediated searches should not only try to assess the validity of the interaction and mathematical models that we propose, by comparing them with unmediated searches, but also try to compare it with alternative or 'competing' approaches. For example, an interaction based on traditional relevance feedback and traditional query expansion techniques such as Rocchio's, could be simulated and compared to our approach.

[1] www.vivisimo.com

7.3.4 User models

We have concentrated on building models for topics in which the user is interested during a search session. We intend to look at recording and combining topic models over search sessions, in order to build user profiles. This will allow our mediation model to be extended in several directions:

- recognise and combine common or similar topics.
- model user profiles as sets of distinct topics.
- model the user's change of interest over time.

7.3.5 Real user experiments

The mediated search simulations have been an excellent tool for understanding the effect of various parameters on the effectiveness of mediation and for comparing various mediation strategies. However, real user experiments are necessary in order to validate our ideas. Our intention is not only to compare mediated and un-mediated searches, but also to compare system-mediated with searches mediated by human search intermediaries. For this endeavour, the mediation system will have to be extended in order to offer functionality comparable to that offered by the human mediator[2]:

- Allow the users to express their interest in natural language.
- Rely on user profiles, learnt during past search sessions.
- Offer a choice of source collection, as well as descriptions and recommendations.
- Offer the user a review of the search session, based on the history of the queries used, the sub-topics explored, and other actions taken during the interaction.

While, in our opinion, we should start with more simple, controlled user experiments, in order to understand the contribution of various formulae and parameters, a more complex and primarily quantitative experiment will be needed to indicate if a system can really replace the human mediator.

[2] Judit Santon, SCILS, Rutgers University, personal discussion.

Bibliography

[All95] James Allan. Relevance feedback with too much data. In *Proceedings of SIGIR'95*, pages 337–343, Seattle, 1995. ACM.

[APC01a] J. A. Anderson and J. Perez-Carballo. The nature of indexing: How humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing and Management*, 37:231–254, 2001.

[APC01b] J. A. Anderson and J. Perez-Carballo. The nature of indexing: How humans and machines analyze messages and texts for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort. *Information Processing and Management*, 37:255–277, 2001.

[AS94] C. Ahlberg and B. Shneiderman. Visual information seeking using filmfinder. In *Proceedings of CHI'94*, pages 433–434, New York, April 1994. ACM. Video.

[Bat89] M. J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424, 1989.

[Bat90] Marcia J. Bates. Where should the person stop and the information search interface start? *Information Processing and Management*, 26(5):575–591, 1990.

[BBDHB94] N. Belkin, C. L. Borgman, S. Dumais, and M. Hancock-Beaulieu. Panel: Evaluating interactive retrieval systems. In *Proceedings of SIGIR'94*, page 361, Dublin, July 1994. ACM.

[BBMS98] M. Barja, T. Bratvold, J. Myllymaeki, and G. Sonnenberger. Informia: a mediator for integrated access to heterogeneous information sources. In *Proceedings of CIKM '98*, pages 234–241, Bethesda, November 1998.

[BDPJ97] M. Beaulieu, T. Do, A. Payne, and S. Jones. ENQUIRE Okapi project. British Library Research and Innovation Report 17, Centre for Interactive Systems Research, City University, London, January 1997.

[Bel93] Nicholas J. Belkin. Interaction with texts: Information retrieval as information-seeking behavior. In *Information Retrieval '93. Von der Modellierung zur Anwendung.*, pages 55–66. Universitaetsverlag Konstanz, 1993.

[Bel98] Nicholas J. Belkin. An overview of results from Rutgers' investigations of interactive information retrieval. In *Visualizing Subject Access for 21st Century Information Resources. Proceedings of the 34th Annual Clinic on Library Applications of Data Processing*, Urbana-Champaign, 1998.

[Bel00] Richard K. Belew. *Finding Out About - A Cognitive Perspective on Search Engine Technology and the WWW.* Cambridge University Press, 2000.

[BI97] P. Borlund and P. Ingwersen. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3):225–250, June 1997.

[BI99] P. Borlund and P. Ingwersen. The application of work tasks in connection with the evaluation of interactive information retrieval systems: Empirical results. Mira Workshop, Glasgow, April 1999.

[BJ94] Kent Beck and Ralph Johnson. Patterns generate architectures. In *European Conference on Object-Oriented Programming*, pages 21–35, Bologna, Italy, July 1994.

[BK02] D. C. Blair and S. O. Kimbrough. Exemplary documents: a foundation for information retrieval design. *Information Processing and Management*, 38(3):363–379, May 2002.

[BMC93] N. J. Belkin, P. G. Marchetti, and C. Cool. Braque: Design of an interface to support user interaction in information retrieval. *Information Processing and Management*, 29(3):325–344, 1993.

[BOB82] N. J. Belkin, R. N. Oddy, and H. M. Brooks. ASK for information retrieval: Part I. Background and theory. *Journal of Documentation*, 38(2):61–71, June 1982.

[BP74] Julie Bichteler and Ronald G. Parsons. Document retrieval by means of an automatic classification algorithm for citations. *Information Storage and Retrieval*, 10:267–278, 1974.

[BRR96] Micheline Beaulieu, Stephen Robertson, and Edie Rasmussen. Evaluating interactive systems in TREC. *Journal of the American Society for Information Science*, 47(1):85–94, 1996.

[BSR93] Peter Bollmann-Sdorra and Vijay V. Raghavan. On the dellusiveness of adopting a common space for modeling IR objects: Are queries documents ? *Journal of the American Society for Information Science*, 44(10):579–587, 1993.

[Bur95] R. Burgin. The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity. *Journal of the American Society for Information Science*, 46(8):562–572, 1995.

[BYRN99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, editors. *Modern Information Retrieval.* ACM Press and Addison-Wesley, 1999.

[Cam95] Iain Campbell. Supporting information needs by ostensive definition in an adaptive information space. In *MIRO'95*, http://www.ewic.org.uk/ewic/, 1995. electronic Workshops in Computing, Springer Verlag.

[Can93] Fazli Can. Incremental clustering for dynamic information processing. *ACM Transactions on Information Systems*, 11(2):143–164, April 1993.

[CCH92] J. P. Callan, W. B. Croft, and S. M. Harding. The INQUERY retrieval system. In *DEXA'92*, pages 78–83, Valencia, Spain, 1992. Springer-Verlag.

[CCL01] J. Callan, W. B. Croft, and J. Lafferty, editors. *Workshop on Language Modelling and Information Retrieval*, Carnegie-Mellon University, May-June 2001. http://la.lti.cs.cmu.edu/callan/Workshops/lmir01/.

[CH79] W. Bruce Croft and David J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35(4):285–295, December 1979.

[Cha00] Rachel Chalmers. Surf like a bushman. *New Scientist*, November 2000.

[CKPW92] D. R. Cutting, D. R. Karger, J. O. Pedersen, and Tukey J. W. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of SIGIR'92*, pages 318–329, Copenhagen, 1992. ACM.

[CLRC98] F. Crestani, M. Lalmas, C. J. van Rijsbergen, and I. Campbell. "Is this document relevant? ... probably". A survey of probabilistic models in information retrieval. *ACM Computing Surveys*, 30(4), 1998.

[CO83] Fazli Can and Esen A. Ozkarahan. A clustering scheme. In *The 6th Annual International ACM-SIGIR Conference*, pages 115–121, New-York, June 1983.

[CO85] F. Can and E. A. Ozkarahan. Concepts of the cover coefficient-based clustering methodology. In *Proceedings of SIGIR'85*, pages 204 –211, Montreal, June 1985. ACM.

[CO90] Fazli Can and Esen A. Ozkarahan. Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. *ACM Transactions on Database Systems*, 15(4):483–517, December 1990.

[CR96] I. Campbell and C. J. van Rijsbergen. The ostensive model of developing information needs. In Peter Ingwersen, editor, *Proceedings of Conceptions of Library and Information Science (CoLIS-2)*, Copenhagen, October 1996.

[Cra92] Walt Crawford. Starting over: Current issues in online catalog user interface design. *Information Technology and Libraries*, 11(1):62–76, March 1992.

[Cro77] W. Bruce Croft. Clustering large files of documents using the single-link method. *Journal of the American Society for Information Science*, 28(6):341–344, November 1977.

[Cro78] William Bruce Croft. *Organizing and Searching Large Files of Document Descriptions.* PhD thesis, Churchill College, Cambridge, October 1978.

[Cro80] W. Bruce Croft. A model of cluster searching based on classification. *Information Systems*, 5:189–195, 1980.

[Cro95] W. Bruce Croft. What do people want from information retrieval ? *D-Lib Magazine*, November 1995.

[CS95] J. O. Coplien and D. C. Schmidt, editors. *Pattern Languages of Program Design.* Addison-Wesley, Reading, 1995. ISBN 0-201-60734-4.

[DDF+90] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[DFK88] S. T. Dumais, G. W. Furnas, and Landauer T. K. Using latent semantic analysis to improve access to textual information. In *Proceedings of CHI'88*, pages 281–285. ACM, 1988.

[DJ79] Richard Dubes and Anil K. Jain. Validity studies in clustering methodologies. *Pattern Recognition*, 11:235–254, 1979.

[DT88] W. J. Doll and G. Torkzadeh. The measurement of end-user computing satisfaction. *MIS Quarterly*, 12:259–274, June 1988.

[Dub95] David Dubin. Document analysis for visualization. In *Proceedings of SIGIR '95*, pages 199–204, Seattle, 1995. ACM.

[Dub96] David Dubin. *Structure in Document Browsing Spaces.* PhD thesis, School of Information Sciences, University of Pittsburgh, 1996.

[Dun97] Mark Dunlop. The effect of accessing non-matching documents on relevance feedback. *ACM Transactions on Information Systems*, 15(2), 1997.

[EFHW93] D. Ellis, J. Furner-Hines, and P. Willett. Measuring the degree of similarity between objects in text retrieval systems. *Perspectives in Information Management*, 3(2):128–149, 1993.

[Eft00] Efthimis N. Efthimiadis. Interactive query expansion: a user-based evaluation in a relevance feedback environment. *JASIS*, 51(11):989–1003, 2000.

[EHW87] Abdelmoula El-Hamdouchi and Peter Willett. Techniques for the measurement of clustering tendency in document retrieval systems. *Journal of Information Science*, 13:361–365, 1987.

[EHW89] Abdelmoula El-Hamdouchi and Peter Willett. Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal*, 32(3):220–227, 1989.

[FBY92] William B. Frakes and Ricardo Baeza-Yates, editors. *Information Retrieval - Data Structures and Algorithms.* Prentice Hall, Englewood Cliffs, New Jersey, 1992.

[FHH96] J. Furner, D. J. Harper, and D. G. Hendry. Coordinated support for browsing, searching and monitoring: A user interface for networked information retrieval. IR/HCI workshop in Glasgow, September 1996.

[Fol96] Peter W. Foltz. Latent semantic analysis for text-based research. *Behaviour Research Methods, Instruments and Computers*, 28(2):197–202, 1996.

[Gea99] David Geary. *Graphic Java - Mastering the JFC*, volume 2 - Swing. Sun Microsystems Press, Palo Alto, 3rd edition, 1999. ISBN 0-13-079667-0.

[GH01] A. Goker and D. He. Personalization in Web retrieval systems: A cross-session approach originated in traditional search environments. In *IJCAI-01 Workshop on Intelligent Techniques for Web*, February 2001.

[GHJV95] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software.* Addison-Wesley, 1995.

[GLW86] Alan Griffiths, H. Claire Luckhurst, and Peter Willett. Using interdocument similarity information in document retrieval systems. *Journal of the American Society for Information Science*, 37(1):3–11, 1986.

[GPS98] G. Golovchinsky, M. N. Price, and B. N. Schilit. Emphasis on the relevant: Free-form digital ink as a mechanism for relevance feedback. In *SIGIR'98*

Workshop on Query Input and User Expectations, Melbourne, August 1998. ACM.

[Har80] David John Harper. *Relevance Feedback in Document Systems: an Evaluation of Probabilistic Strategies.* PhD thesis, Jesus College, Cambridge, February 1980.

[Har92] Donna Harman. Relevance feedback revisited. In *Proceedings of SIGIR'92*, pages 1–15, Copenhagen, Denmark, 1992. ACM.

[Har93] Donna Harman. Overview of the first TREC conference. In *Proceedings of SIGIR '93*, pages 36–47, Pittsburgh, 1993. ACM.

[Hea95] M. A. Hearst. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of CHI'95*, Denver, May 1995.

[Hea97] M. Hearst. Cat-a-Cone: An interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In *Proceedings of SIGIR '97*, pages 246–255, Philadelphia, July 1997. ACM.

[Hea99a] M. A. Hearst. *Natural Language Information Retrieval*, chapter The Use of Categories and Clusters for Organizing Retrieval Results, pages 333–373. Kluwer Academic Publishers, 1999.

[Hea99b] Marti Hearst. *Modern Information Retrieval*, chapter User Interface and Visualization, pages 257–325. ACM Press and Addison-Wesley, 1999.

[Hen96] D. G. Hendry. *Extensible Information-Seeking Environments.* PhD thesis, School of Computer and Mathematical Sciences, The Robert Gordon University, Aberdeen, September 1996.

[HF99] K. Hornbaek and E. Frokjaer. Do thematic maps improve information retrieval ? In M. A. Sasse and C. Johnson, editors, *Interact'99*, pages 179–186. IFIP, IOS Press, 1999.

[HGH01] D. He, A. Goker, and D. J. Harper. Combining evidence for automatic Web session identification. *Information Processing and Management*, 2001. Special issue: Issues of contexts in IR.

[HH96] D. G. Hendry and D. J. Harper. An architecture for implementing extensible information-seeking environments. In *Proceedings of SIGIR '96*, pages 94–100, Zurich, August 1996. ACM.

[HH97] D. G. Hendry and D. J. Harper. An informal information-seeking environment. *JASIS*, 48(11):1036–1048, November 1997.

[Hie01] Djoerd Hiemstra. *Using Language Models for Information Retrieval.* PhD thesis, University of Twente, The Netherlands, January 2001.

[HLH94] W. Hersh, T. J. Leone, and D. Hickam. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of SIGIR'94*, pages 192–201, Dublin, July 1994. ACM.

[HMM99] D. J. Harper, M. Mechkour, and G. Muresan. Document clustering for mediated information access. In *Proceedings of the 21st Annual BCS-IRSG Colloquium*, Glasgow, April 1999.

[HO99] W. Hersh and P. Over. TREC-8 interactive track. *SIGIR Forum*, 33(2):8–11, Winter 1999.

[HP96] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In H.-P. Frei, D. Harman, P. Schauble, and R. Wilkinson, editors, *Proceedings of SIGIR '96*, pages 76–84, Zurich, Switzerland, August 1996. ACM.

[HPP+96] M. Hearst, J. Pedersen, P. Pirolli, H. Schutze, Grefenstette, and D. Hull. Xerox site report: Four TREC-4 tracks. In *Proceedings of TREC-4*, November 1996.

[HR97] William Hersh and Stephen Robertson. Evaluation of information retrieval systems. SIGIR'97 Tutorial, July 1997. Philadelphia.

[Hul93] David Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of SIGIR'93*, pages 329–338, Pittsburgh, USA, 1993. ACM.

[HW80] Alan F. Harding and Peter Willett. Indexing exhaustivity and the computation of similarity matrices. *Journal of the American Society for Information Science*, 31:298–300, 1980.

[HZ80] Karen A. Hamill and Antonio Zamora. The use of titles for automatic document classification. *Journal of the American Society for Information Science*, 31:396–402, 1980.

[JAS98] B. J. Jansen, Spink A., and T. Saracevic. Failure analysis in query construction: Data and analysis from a large sample of Web queries. In *Proceedings of Digital Libraries '98*, pages 298–299, Pittsburgh, 1998. ACM.

[JD88] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, 1988.

[JIH96] Joemon J. Jose, Jan J. IJdens, and David J. Harper. Using Dempster-Shafer theory of evidence in visual information retrieval. Technical report, School of Computer and Mathematical Sciences, The Robert Gordon University, Aberdeen, Scotland, August 1996.

[Jos98] Joemon Jose. *An Integrated Approach for Multimedia Information Retrieval*. PhD thesis, School of Computer and Mathematical Sciences, Robert Gordon University, Aberdeen, Scotland, United Kingdom, April 1998.

[JR71] N. Jardine and C. J. van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7:217–240, 1971.

[JSBS98] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the Web. *SIGIR Forum*, 32(1):5–17, 1998.

[JSS00] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the Web. *Information Processing and Management*, 36:207–227, 2000.

[KB96] Jurgen Koenemann and Nicholas J. Belkin. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proceedings of CHI '96*, pages 205–212, New York, 1996. ACM.

[KB01] D. Kelly and N. J. Belkin. Reading time, scrolling amd interaction: Exploring implicit sources of user preferences for relevance feedback. In *Proccedings of SIGIR'01*, pages 408–409, New Orleans, USA, 2001. ACM.

[Kir95] David Kirsh. The intelligent use of space. *Artificial Intelligence*, 73:31–68, 1995.

[KM00] G. J. Kowalski and M. T. Maybury. *Information Storage and Retrieval.* Kluwer Academic Publishers, Boston, 2nd edition, 2000.

[Kor91] R. R. Korfhage. To see, or not to see - is that the query? In *Proceedings of SIGIR '91*, pages 134–141, Chicago, October 1991. ACM.

[Kor97] R. Korfhage, Robert. *Information Storage and Retrieval.* John Wiley & Sons, New York, 1997.

[KRJ01] Y. Kural, S. Robertson, and S. Jones. Deciphering cluster representations. *Information Processing and Management*, 37:593–601, 2001.

[Kur99] S. Yasemin Kural. *Clustering Information Retrieval Search Outputs.* PhD thesis, Department of Information Science, City University, London, September 1999.

[Kwo75] K. L. Kwok. The use of title and cited titles as document representation for automatic classification. *Information Processing and Management*, 11:201–206, 1975.

[LC01] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proceedings of SIGIR'01*, New Orleans, 2001. ACM.

[LO98] E. Lagergren and P. Over. Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment. In *Proceedings of SIGIR'98*, pages 164–172, Melbourne, 1998. ACM.

[LRP95] J. Lamping, R. Rao, and P. Pirolli. A focus + context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proceedings of CHI'95*, pages 401–408, Denver, Colorado, May 1995. ACM.

[Mar95] G. Marchionini. *Information Seeking in Electronic Environments.* Cambridge University Press, 1995.

[MDKL93] G. Marchionini, S. Dwiggins, A. Katz, and X. Lin. Information seeking in full-text end-user-oriented search systems: The roles of domain and search expertise. *Library and Information Science Research*, 15:35–69, 1993.

[Miz96] Stefano Mizzaro. How many relevances in IR ? Technical Report GR96-2, Computer Science Department, Glasgow University, September 1996.

[Miz97] Stefano Mizzaro. Relevance: The whole history. *JASIS*, 48(9):810–832, September 1997.

[MJS+97] Y. S. Maarek, M. Jacovi, M. Shtalhaim, S. Ur, D. Zernik, and I. Z. Ben Shaul. WebCutter: A system for dynamic and tailorable site mapping. In *The 6th WWW Conference*, pages 713–722, Santa Clara, CA, USA, 1997.

[MM72] Daniel McClure Murray. *Document Retrieval Based on Clustered Files.* PhD thesis, Department of Computer Science, Cornell University, Ithaca, New York, June 1972.

[MR97] M. Magennis and C. J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *Proceedings of SIGIR '97*, pages 324–332, Philadelphia, July 1997. ACM.

[MRB98] R. C. Martin, D. Riehle, and F. Buschmann, editors. *Pattern Languages of Program Design 3.* Addison-Wesley, Reading, 1998. ISBN 0-201-31011-2.

[MS99] Christopher D. Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, Massachusetts, 1999.

[MSS83] G. M. Milligan, S. C. Soon, and L. M. Sokol. The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(1):40–47, 1983.

[MZ97] A. Moffat and J. Zobel. Efficient information retrieval. Tutorial at SIGIR'97, July 1997.

[Nie93] Jakob Nielsen. *Usability Engineering.* Academic Press, 1993.

[Nor88] Donald A. Norman. *The Psychology of Everyday Things.* Basic Books, 1988.

[Nor96] Ragnar Nordlie. Unmediated and mediated information searching in the public library. In *Proceedings of ASIS'96*, 1996.

[Nor99] Ragnar Nordlie. "User revealment" - a comparison of initial queries and ensuing question development in online searching and in human reference interactions. In *Proceedings of SIGIR'99*, pages 11–18, Berkeley, August 1999. ACM.

[NSKF00] J. L. Neto, A. D. Santos, C. A. A. Kaestner, and A. A. Freitas. Document clustering and text summarization. In *Proceedings of PADD 2000*, pages 41–55, London, 2000.

[Odd77] R. N. Oddy. Information retrieval through man-machine dialogue. *Journal of Documentation*, 33(1):1–14, March 1977.

[OJ93] Vicky L. O'Day and Robin Jeffries. Orienteering in an information landscape: How information seekers get from here to there. In *Proceedings of INTERCHI '93*, pages 438–445, Amsterdam, April 1993. ACM.

[Ove01] Paul Over. The TREC interactive track: An annotated bibliography. *Information Processing and Management*, 37:369–381, 2001.

[Par97] Hongseok Park. Relevance of science information: Origins and dimensions of relevance and their implications to information retrieval. *Information Processing and Management*, 33(3):339–352, 1997.

[PC95] Peter Pirolli and Stuart Card. Information foraging in information access environments. In *Proceedings of CHI '95*, pages 51–58, Denver, May 1995. ACM.

[PC98] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of SIGIR '98*, pages 275–281, Melbourne, 1998. ACM.

[PCS00] J. Perez-Carballo and T. Strzalkowski. Natural language information retrieval: Progress report. *Information Processing and Management*, 36:155–178, 2000.

[PCW01] P. Pirolli, S. K. Card, and Mija M. van der Wege. Visual information foraging in a focus+context visualization. In *Proceedings of CHI 2001*, Seattle, 2001. ACM.

[PHF99] W. Pratt, M. Hearst, and L. Fagan. A knowledge-based approach to organizing retrieved documents. In *Proceedings of AAAI-99*, July 1999.

[Pol97a] Steven Pollitt. Interactive information retrieval based on faceted classification using views. In *Knowledge Organization for Information Retrieval, Proceedings of the 6th International Study Conference on Classification*, University College, London, June 1997.

[Pol97b] Steven Pollitt. The key role of classification and indexing in view-based searching. In *Proceedings of IFLA '97*, Copenhagen, 1997.

[Pon00] Jay M. Ponte. *Advances in Information Retrieval*, chapter Language Models for Relevance Feedback, pages 73–96. Kluwer Academic Publishers, 2000.

[Pra99] Wanda Pratt. *Dynamic Categorization: A Method for Decreasing Information Overload.* PhD thesis, Stanford Medical Informatics, Stanford University, March 1999.

[PRSB94] J. Preece, Y. Rogers, H. Sharp, and D. Benyon. *Human-Computer Interaction.* Addison-Wesley, 1994. ISBN 0-201-62769-8.

[PSHD96] P. Pirolli, P. Schank, M. Hearst, and C. Diehl. Scatter/Gather browsing communicates the topic structure of a very large text collection. In *Proceedings of CHI '96*, Vancouver, Canada, April 1996. ACM.

[QF93] Y. Qiu and H. P. Frei. Concept based query expansion. In *Proceedings of SIGIR'93*, pages 160–169, Pittsburgh, USA, 1993. ACM.

[Ras92] Edie Rasmussen. *Information Retrieval - Data Structures and Algorithms*, chapter Clustering Algorithms, pages 419–442. Prentice Hall, Englewood Cliffs, New Jersey, 1992.

[RC75] C. J. van Rijsbergen and W. B. Croft. Document clustering: An evaluation of some experiments with the Cranfield 1400 collection. *Information Processing and Management*, 11:171–182, 1975.

[Rei00] Jane Reid. A task-oriented non-interactive evaluation methodology for information retrieval systems. *Information Retrieval*, 2(1):113–127, 2000.

[RH01] S. Robertson and D. Hiemstra. Language models and probability of relevance. In *Workshop on Language Modelling and Information Retrieval*, Carnegie-Mellon University, May-June 2001.

[RHB92] S. E. Robertson and M. M. Hancock-Beaulieu. On the evaluation of IR systems. *Information Processing and Management*, 28(4):457–466, 1992.

[Rij79] C. J. van Rijsbergen. *Information Retrieval.* Butterworth and Co, London, 2nd edition, 1979.

[RMC91] G. G. Robertson, J. D. Mackinlay, and Stuart K. Card. Cone Trees: Animated 3D visualizations of hierarchical information. In *Proceedings of CHI '91*, pages 189–194, New-Orleans, 1991. ACM.

[Rob77] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, December 1977.

[RPG94] J. Rasmussen, A. M. Pejtersen, and L. P. Goodstein. *Cognitive Systems Engineering.* John Wiley & Sons, New York, 1994.

[RPH+95] R. Rao, J. O. Pedersen, M. A. Hearst, J. D. Mackinlay, S. K. Card, L. Masinter, P.-K. Halvorsen, and G. G. Robertson. Rich interaction in the digital library. *Communications of the ACM*, 38(4):29–39, April 1995.

[RSJ73] C. J. van Rijsbergen and Karen Sparck Jones. A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29(3):251–257, 1973.

[RSJ97] S. E. Robertson and K. Sparck Jone. Simple, proven approaches to text retrieval. Technical Report 356, Computer Laboratory, Cambridge University, May 1997.

[RWB00] S. E. Robertson, S. Walker, and M. Beaulieu. Experimentation as a way of life: Okapi at TREC. *Information Processing and Management*, 36:95–108, 2000.

[Sal68] Gerard Salton. *Automatic Information Organization and Retrieval.* McGraw-Hill, New York, 1968.

[Sar95] T. Saracevic. Evaluation of evaluation in information retrieval. In *Proceedings of SIGIR '95*, pages 138–146. ACM, July 1995.

[SB90] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *JASIS*, 41(4):288–297, 1990.

[SBCQ98] A. F. Smeaton, M. Burnett, F. Crimmins, and G. Quinn. An architecture for efficient document clustering and retrieval on a dynamic collection of newspaper texts. In Keith van Rijsbergen, editor, *Proceedings of BCS-IRSG '98*, Grenoble, March 1998.

[SC99] F. Song and W. B. Croft. A general language model for information retrieval. In *Proceedings of CIKM '99*, Kansas City, August 1999.

[SGR98] A. Spink, A. Goodrum, and D. Robins. Elicitation behaviour during mediated information retrieval. *Information Processing and Management*, 34(2/3):257–273, 1998.

[SGRM96] A. Spink, A. Goodrum, D. Robins, and Wu M. M. Elicitations during information retrieval: Implications for IR system design. In *Proceedings of SIGIR'96*, pages 120–127, Zurich, Switzerland, 1996. ACM.

[Shn92] Ben Shneiderman. Tree visualization with Tree-Maps: 2-D space-filling approach. *ACM Transactions on Graphics*, 11:92–99, January 1992.

[Shn98] Ben Shneiderman. *Designing the User Interface. Strategies for Effective Human-Computer Interaction.* Addison-Wesley, third edition, 1998.

[SJ71] Karen Sparck Jones. *Automatic Keyword Classification for Information Retrieval.* Butterworths, London, 1971.

[SJ73] Karen Sparck Jones. Collection properties influencing automatic term classification performance. *Information Storage and Retrieval*, 9:499–513, 1973.

[SJ00] K. Sparck Jones. Further reflections on TREC. *Information Processing & Management*, 36(1):37–85, January 2000.

[SJBH97] W. M. Shaw Jr, R. Burgin, and P. Howell. Performance standards and evaluations in IR test collections: Cluster-based retrieval models. *Information Processing and Management*, 33(1):1–14, January 1997.

[SJN68] K. Sparck Jones and R. M. Needham. Automatic term classification and retrieval. *Information Storage and Retrieval*, 4:91–100, 1968.

[SKK00] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *TextMining Workshop, KDD2000*, 2000.

[SM83] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* Computer Science. McGraw-Hill, Inc, New-York, 1983.

[Spe00] Robert Spence. *Information Visualization.* Longman, 2000. ISBN 0-201-59626-1.

[Spi97] Amanda Spink. Study of interactive feedback during mediated information retrieval. *JASIS*, 48(5):382–394, 1997.

[Spo94] Anselm Spoerri. InfoCrystal: Integrating exact and partial matching approaches through visualization. In *Proceedings of RIAO '94*, pages 687–696, October 1994.

[SS73] Peter H. A. Sneath and Robert R. Sokal. *Numerical Taxonomy.* W. H. Freeman and Company, San Francisco, 1973.

[SS85] H. Small and E. Sweeney. Clustering the science citation index using cocitations. *Scientometrics*, 7(3-6):391–409, 1985.

[SSW97] T. Saracevic, A. Spink, and Mei-Mei Wu. Users and intermediaries in information retrieval: What are they talking about ? In A. Jameson, C. Paris, and C. Tasso, editors, *Proceedings of UM '97*, Vienna, 1997. Springer.

[SW78] G. Salton and A. Wong. Generation and search of clustered files. *ACM Transactions on Database Systems*, 3(4):321–346, December 1978.

[SWY75] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[Tay68] Robert S. Taylor. Question-negotiation and information seeking in libraries. *College and Research Libraries*, 29:178–194, 1968.

[VCK96] J. M. Vlissides, J. O. Coplien, and N. L. Kerth, editors. *Pattern Languages of Program Design 2.* Addison-Wesley, Reading, 1996. ISBN 0-201-89527-7.

[VH00] E. M. Voorhees and D. Harman. Overview of the sixth Text REtrieval Conference. *Information Processing and Management*, 36:3–35, 2000.

[Voo85a] E. M. Voorhees. The cluster hypothesis revisited. Technical Report 85-658, Department of Computer Science, Cornell University, Ithaca, New York, April 1985.

[Voo85b] E. M. Voorhees. *The Effectiveness and Efficiency of Agglomerative Hierarchic Clustering in Document Retrieval.* PhD thesis, Department of Computer Science, Cornell University, Ithaca, October 1985.

[Voo86] E. M. Voorhees. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing and Management*, 22(6):465–476, 1986.

[Wil80] P. Willett. Document clustering using an inverted file approach. *Journal of Information Science*, 2:223–231, 1980.

[Wil81] P. Willett. A fast procedure for the calculation of similarity coefficients in automatic classification. *Information Processing and Management*, 17:53–60, 1981.

[Wil83] P. Willett. Similarity coefficients and weighting functions for automatic document classification: an empirical comparison. *International Classification*, 10(3):138–142, 1983.

[Wil84] P. Willett. A note on the use of nearest neighbors for implementing single linkage document classifications. *Journal of The American Society for Information Science*, 35(3):149–153, 1984.

[Wil85] P. Willett. Query-specific automatic data classification. *International Forum on Information on Documentation*, 10(2):28–32, April 1985.

[Wil88] P. Willett. Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management*, 24(5):577–597, 1988.

[Wil96] Peter Willett. Document clustering: The myth and the reality. Talk presented at the Robert Gordon University, Aberdeen, June 1996.

[WJR01] R. W. White, J. M. Jose, and I. Ruthven. Comparing explicit and implicit feedback techniques for web retrieval: Trec-10 interactive track report. In E. M. Voorhees and D. K. Harman, editors, *The Tenth Text REtrieval Conference (TREC 2001)*. NIST, 2001.

[WMB99] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes. Compressing and Indexing Documents and Images*. Morgan Kaufmann, San Francisco, 2nd edition, 1999. ISBN 1-55860-570-3.

[Wor69] S. Worona. *Scientific Report No. ISR-16 - Information Storage and Retrieval - to the National Science Foundation*, chapter Query Clustering in a Large Document Space, pages XV–1–XV–22. Department of Computer Science, Cornell University, Ithaca, USA, September 1969.

[WS92] C. Williamson and B. Shneiderman. The dynamic HomeFinder: Evaluating dynamic queries in a real-estate information exploration system. In *Proceedings of CHI'92*, pages 338–346, Copenhagen, 1992. ACM.

[XC96] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proceedings of SIGIR '96*, pages 4–11, Zurich, 1996. ACM.

[XC99] J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *Proceedings of SIGIR '99*, pages 254–261, Berkeley, August 1999. ACM.

[XC00] J. Xu and W. B. Croft. *Advances in Information Retrieval*, chapter Topic-Based Language Models for Distributed Retrieval, pages 151–172. Kluwer Academic Publishers, 2000.

[YL77] C. T. Yu and W. S. Luk. Analysis of effectiveness of retrieval in clustered files. *Journal of the ACM*, 24(4):607–622, October 1977.

[ZE98] O. Zamir and O. Etzioni. Web document clustering: A feasibility demonstration. In *Proceedings of SIGIR '98*, pages 46–54, Melbourne, August 1998. ACM.

[ZE99] O. Zamir and O. Etzioni. Grouper: A dynamic clustering interface to Web search results. In *Proceedings of WWW8*, 1999.

[ZEMK97] O. Zamir, O. Etzioni, O. Madani, and R. M. Karp. Fast and intuitive clustering of Web documents. In *The Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*. AAAI, 1997.

Published Papers

- Gheorghe Muresan and David J. Harper(2001). Document Clustering and Language Models for System-Mediated Information Access. *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'01)*, Darmstadt, September 2001, pp.438-449, ISBN 3-540-42537-3.

- Gheorghe Muresan, David J. Harper and Ayse Goker(2001). ClusterBook, a Tool for System-Mediated Access via Clustered Collections. *ECDL'01*, Darmstadt, Germany, September 2001 (Demo/poster).

- Gheorghe Muresan, David J. Harper, Ayse Goker and Peter Lowit(2000). ClusterBook, a Tool for Dual Information Access. *Proceedings of SIGIR'00*, Athens, July 2000, pp.391 (Demo).

- Gheorghe Muresan, David J. Harper and Mourad Mechkour(1999). WebCluster, a tool for Mediated Information Access. *Proceedings of SIGIR'99*, Berkeley, August 1999, pp.337 (Demo).

- David J. Harper, Mourad Mechkour and Gheorghe Muresan(1999). Document Clustering for Mediated Information Access. *Proceedings of the 21st BCS-IRSG Annual Colloquium on IR Research*, Glasgow, Scotland, April 1999.

- Mourad Mechkour, David J. Harper and Gheorghe Muresan (1998). The WebCluster Project. Using Document Clustering for Mediating Access to the World Wide Web. *Proceedings of SIGIR'98*, Melbourne, Australia, August 1998 (Poster).