Structured text retrieval by means of affordances and genre

Malcolm Clark
School of Computing
The Robert Gordon University
mc@comp.rgu.ac.uk

ABSTRACT

This paper offers a proposal for some preliminary research on the retrieval of structured text, such as extensible mark-up language (XML). We believe that capturing the way in which a reader perceives the meaning of documents, especially genres of text, may have implications for information retrieval (IR) and in particular, for cognitive IR and relevance. Previous research on 'shallow' features of structured text has shown that categorization by form is possible. Gibson's theory of 'affordances' and genre offer the reader the meaning and purpose - through structure - of a text, before the reader has even begun to read it, and should therefore provide a good basis for the 'deep' skimming and categorization of texts. We believe that Gibson's 'affordances' will aid the user to locate, examine and utilize shallow or deep features of genres and retrieve relevant output. Our proposal puts forward two hypotheses, with a list of research questions to test them, and culminates in experiments involving the studies of human categorization behaviour when viewing the structures of emails and web documents. Finally, we will examine the effectiveness of adding structural layout cues to a Yahoo discussion forum (currently only a bag-of-words), which is rich in structure, but only searchable through a Boolean search engine.

Keywords: categorization, genre, affordances, skimming.

1. INTRODUCTION

IR overlaps with numerous other fields of research: artificial intelligence (AI), human computer interaction (HCI) and natural language processing (NLP), to name but a few. This paper, however, focuses on overlaps of visual perception, genre and AI, merging and utilizing these for one particular goal: structured information text retrieval through skimming and categorization.

The difficulties inherent in IR are many and various, but the over-riding problem has been how to find and display results with the highest relevance to users' needs, whether these are in the form of video, text, audio files or images. As a result, searching has been perceived as the most notable aspect of IR and has, up to now, been given the most recognition. Research involving the retrieval of structured text or document retrieval, however, is currently progressing at a rapid pace [1, 2].

As the quantity and size of the XML or extensible hypertext mark-up language (XHTML) document collections continue to expand (as web and digital libraries, for example) the need for IR systems which exploit structured text, as opposed to traditional bag-of-words (B-O-W) based IR systems, is also increasing. Structured textual documents are normally composed of several layers or sections which together form genres of text preserved, in particular, in XML/XHTML. A genre in this context is the set of structures, layout and style of writing (or as Dewdney et al.[3] state "conventions") which show the user the documents' purpose and form through its structure regardless of the topical nature of the writing. Since XML retains genre information, such documents should be explored for genre.

Genres have been debated for thousands of years, see, for example, most notably, Plato, with his "*Theory of Forms*" [4]. Genre has been cited in relation to the European Romantic movement of the 18th and 19th centuries, the Russian Formalists [5] and, arguably, most famously by Bahktin [6] in his essays on 'Speech Genres'.

Genres have been used to categorize many things, such as music, prog-rock and punk; literary works, tragedy, comedy, etc, and as Yates et al. [7] explained, organizational communication: "In structurational terms, genres are social institutions that are produced, reproduced, or modified when human agents draw on genre rules to engage in organizational communication".

Our current work [8], which originally focused on the shallow features of genre, has now been extended to analyse the identification processes and actions employed by readers when searching for a relevant text, especially when skimming. Following Gibson's theory of 'affordances' [9], we examine the 'deep' features of genre, which are used by readers to determine the meaning and purpose of texts.

Genre could hold great potential benefits for organisations, both financially and administratively, by allowing automatic and rapid information retrieval without the need for manual organization and sorting. In particular, the

sorting and filtering of emails would benefit large organisations by improving their operational capability and reducing time consuming tasks.
Section 2 describes the background to this research: XML, text categorization, genre, deep and shallow parsing of genre features, visual perception and affordances. Section 3 lists our hypotheses, whilst Section 4 outlines proposed experiments and issues for discussion. The paper closes with conclusions drawn from our research.

2. BACKGROUND

2.1. XML

XML and offshoot XHTML are becoming dominant formats for managing structured text, especially on the world wide web (WWW), and are in widespread use in many fields, varying from digital libraries and electronic publishing to the so called 'semantic web', for example, Wikipedia (XHTML). Although this research is mainly involved in whole document retrieval there may also be benefits for partial section retrieval.
The form or structural information contained within the interface of specific XML or maintained in the tag information needs to be exploited. "The increase in storage of digital documents in XML format has brought about the explosion in the development of systems to store, access and exploit the logical structure of such corpuses." Lalmas et al. [10]
In this project, we investigate the usefulness of form features for the retrieval of structured documents. During earlier experiments, in press [8], useful indicators were identified which demonstrate the effectiveness of using genre for the purpose of text categorization, also known as text classification. We have found that genres can be discriminated by utilizing the shallow features of tags and grammar [11], that is, the structural information (or form) within the XML/XHTML documents.

2.2. Text categorization

The categorisation of documents is normally implemented by labelling and classification: "Text categorization (TC – also known as text classification, or topic spotting) is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefined set." Sebastiani [12]. By grouping collections of documents into smaller groups which are usually pre-labeled, genre is utilized by the form (and sometimes content, style, function etc) of the documents whilst text classification traditionally discriminates by topic using such features as keywords.
The applications of text categorization are numerous: automatic essay grading, spam filtering, and of course, genre filtering. Most categorization of digital media is based around the topical aspect of a document; research into genre therefore needs to diversify. Many authors, such as Rauber et al. [13], Benno Stein and zu Eissen [14], and Sebastiani [12] have recognized the value of genre in conventional libraries and information searches.
Given the increase in diversity of document genres in digital libraries, there is clearly a desperate need for improvements in the organization and management of documents. Most retrieval is based on structure and content in collections, and not solely on structure, but this needs to change as genre types provide excellent features for distinguishing among types of documents.

2.3. Genre

Broadly speaking, genre research can be divided into two main schools of thought [15, 16]: the North American School and the Sydney School. The former views genre as a socio-historical, rhetorically-oriented concept, with the emphasis placed on how texts function in social and interactional contexts [7], while the Sydney School is based on an applied linguistic approach, with the focus on formal textual features. For more information, the reader is directed to Breuer's [17] article.
Contemporarily, a search for 'genre' in dictionaries usually reveals definitions such as: the classification of movies or literature as, for example, 'westerns', 'short stories' or 'detective novels'. Although, of course, this definition has some relevance, the term 'genre' embodies a much wider range of contexts. Watt [18], for example, refers to the opportunities offered by the "socially constructed communicative behaviours called genres to improve the efficiency of communal activities" and Yates et al. [7], in their pioneering work on the concept of genre, suggest that "Genres (e.g., the memo, the proposal, and the meeting) are typified communicative actions characterized by similar substance and form and taken in response to recurrent situations", which can be used to identify types of organizational communication.
Current research on genre is based around three areas of interest [14]: literary theories of genre (including kinds of literature [6], automatic genre identification, and genre and the WWW [19, 20]. Although our research is mostly focused on automatic genre identification and genre and the WWW, the literary genres with the shallow and deep approaches explained in 2.4 and 2.5 are also applicable to genre identification and genre and the WWW.
A search engine using the structural information implicit in most documents would assist the user to retrieve the correct type of information and articles, with high performance and effective filtering of the results. This is the key to genre: it is a way of exploring the standardized types of communication that emerge within a community of practice [21].
There are many diverse types of genres (academic and scientific articles, biographies, news articles, memos, newsletters) and we have noted that new and emergent genres evolve gradually on email exchanges and social networking sites, for example, Wikipedia.
Standard approaches to genre classification are often inadequate because the standard interpretation of user needs only takes into account the need to discriminate by topic, not by genre. When a user inputs a query into a search engine, this usually results in a wide range of documents of different topics being returned which need to be sifted through. For example, if the search query 'Tom Clancy Biography' is entered, the results returned contain interviews, web chats, biographies, book reviews and so on. We employ a genre-based approach, incorporating

the deep and shallow parsing of genre features, to examine the textual features of documents, to categorize the purposes for which the text has been written and to identify the form elements which are common to various genres. We argue that essential improvements in performance can be achieved by using structural information to filter and reduce the cognitive load.

2.4. Shallow parsing – genre retrieval

Numerous shallow feature parsing experiments have been carried out with regard to XML/XHTML retrieval [22, 23], but not on structured text for XML and XHTML genre research. Many papers have, however, been written on web genre research techniques [24, 14, 25, 26] which are applicable for most digital collections.

There are several underlying concepts that are persistent in genre definitions: the style, form, content and functionality of the document. Web genres incorporate the style, form, and content of the document, which are orthogonal and not related to the topic or classification of many genres. These three concepts are probably applied most consistently within WWW IR circles. The conceptual features of style, format, content and functionality can be used with other digital document formats, especially XML.

Campbell & Toms [27] suggest that the conceptual features consist of a grouping of unique facets or levels, i.e. function, form and interface. The function is representative of the meanings of the words contained within the documents, the form refers to the layout or appearance, and, finally, the interface is the way in which the document is read or used. By looking at these conceptual features, many pieces of genre classification work can be seen to fit with the concepts they describe (Table 1).

TABLE 1. Concept examples

Concept	Small Selection of Feature Examples
Style	Readability and part-of-speech (POS) statistics. [24, 28]
Form	Text statistics, whitespace, and formatting tag analysis.[24, 11, 8, 29]
Content	Terms, words in HTML title tag and uniform resource locator (URL), number types, closed-world sets, punctuation. [24, 30]
Functionality	Number of links in a web page; number of e-mail links. [29]

Taking style, form, content and function into consideration, there are hundreds of features that can be measured and the normal practice is to group them as feature sets. There is some debate regarding the overlapping of the sets to which some features can be assigned, but this overlap does serve a genuine purpose. It enables the classification to be tested against each feature set, for example, style versus form. Some documents do, of course, provide visual markers that allow the reader to conceptualize the format. Documents contain distinguishable features such as patterns which allow the reader to identify the purpose and content of a document (figure. 1).

Call for Papers
The First International Workshop on
Data Mining and Audience Intelligence for Advertising (ADKDD'07)
August 12, 2007, San Jose, California, USA
URL: http://adlab.microsoft.com/adkdd2007/
in conjunction with
The 13th International Conference on Knowledge Discovery and Data Mining
(SIGKDD'07)
****

Topics

- Auction theory and design for advertising
- Search intent discovery for advertising
- Audience intelligence

FIGURE 1. A standard (shortened) formatted email 'call for papers'

A good account of the debate is found in Luštrek [31] who defines these features and concepts.

We are proposing a new approach or set of features but this still belongs to the form (or structural) concept. Although shallow parsing methods for automatic genre retrieval are important to the project, our intention is also to extend the deep parsing methods to test whether genre could be useful for skimming texts.

2.5. Deep parsing – skimming for genre

Skimming is a technique for reading which is utilized to identify the main purpose of the text. It is performed at a speed several times faster than conventional reading and is normally used when a reader has a large amount of text to read and does not need to understand every word, for example, when a student has to perform a literature search. This technique dovetails neatly with the theory of affordances and genre. Watt [32] states: "These [genres] are there to reduce cognitive load – there is no need for a person to read the whole text, the genre provides these filtering cues in its structure."

The importance of genre for skimming text has yet to be realised in IR. Skimming has previously been used in the natural language processing (NLP) framework, for example, by DeJong [33] and Mauldin [34]. DeJong developed a natural language parser named fast reading and understanding memory program (FRUMP) and Mauldin extended this parser (McFRUMP) in his flexible expert retrieval of relevant English text (FERRET) system. The skimming works detailed above are all works within narrow domains. Our research targets broader domains in IR – using

emails, technologically structured documents and also documents structured through social consensus. Watt's [35] Open Book and Sentinel applications (web and email respectively) were inspired by De Jong's FRUMP (Predictor/Substantiator).We contend that documents acting as affordances (2.6) can also represent form and meaning (or purpose) to the reader by utilizing the genre rules, structure and patterns.

2.6. Visual perception and affordances

Many approaches have been developed for the study of visual perception, for example, Gestalt Theory, according to which "an image tends to be perceived according to the organization of the elements within it, rather than according to the nature of the individual elements themselves" [36]. The laws of closure, proximity and similarity are important to this theory but a full description is beyond the scope of this review. Another example is provided by Marr, who identified different stages of perception to those identified by Gibson, and, most importantly, put forward the theory that the final stage in perception was recognition and not action. [36]

The bottom-up approach and notion of affordance was first introduced by J.J Gibson [9] to explain his theories of visual perception. He proposed an alternative direct perception framework [37], which cuts against the grain of traditional belief, and his theory of 'affordances', in particular, is still very influential in the field of ecological psychology. Gibson [9] coined the term "affordance" and then explained how an affordance offers an 'animal' a particular actionable possibility which is independent of the animal's ability to visually perceive any possibility. He claimed that we perceive the affordance properties of the environment in a direct and immediate way. He discounted the theories contained within physics and moved towards the ideas of ecological psychology.

To explain affordances, we will briefly examine Gibson's cornerstone work. He defined affordances in the context of visual perception by explaining how animals visualise at the same levels of mediums, surfaces and atoms and perceive what the combinations of the three offer: "…the affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill." [9]

According to Gibson affordances contain the following attributes:

1. Affordances provided by the environment are what it offers, what it provides, what it furnishes, and what it invites.
2. The "values" and "meanings" of things in the environment can be directly perceived. The "values" and "meanings" are external to the perceiver.
3. Affordances are relative to animals. They can only be measured in ecology, not in physics.
4. An affordance is an invariant.
5. Affordances are holistic. What we perceive when we look at objects are their affordances, not their dimensions and properties.
6. An affordance implies the complementarity of the perceiver and the environment. It is neither an objective property nor a subjective property, and at the same time it is both. It cuts across the dichotomy of subjective-objective. Affordances only make sense from a systemic point of view.

Examples of affordances are compiled by Zhang and Patel [38] who have attempted to compile a taxonomy in which they categorize possible affordances.

The difficulties inherent in compiling such a work are obvious: take, for example, a bunch of grapes, which affords so much to different types of animals. Grapes can afford food, pleasure and sustenance to a human or poison and death to canines.

Unfortunately, the examples given by Zhang and Patel regarding cognitive and perceptive affordances are a little confused. The two perceptual examples, "stove-dials and hot-plates" and "men's and women's toilet signs" perhaps belong in a different category, under semiotics.

We argue that a genre 'affords' its meaning and purpose in which the reader perceives the invariants in the ambient optic array and we intend to locate, examine and hopefully utilize the invariant cues (or layout of texts) such as whitespace for the purposes of structured text retrieval.

3. HYPOTHESES

During this first year of research into structured text retrieval, two hypotheses have emerged. The most important of these are:

1. If genre studies and ecological psychology are applied to current information retrieval techniques, they will lead to an improved performance in the retrieval of structured documents.
2. If Gibson's affordances are utilized for object perception and/then meaning, they will be useful for genre and the retrieval of structured texts.

In order to test the validity of the hypotheses mentioned above, we seek to answer the following questions:

1. Can the exploitation of structural information such as tag counts and genre identification be helpful for retrieval?
2. Can documents be categorized or identified without classification and labeling to show what they afford?
3. Does layout assist readers to classify texts?
4. Are layout features independent of linguistic features?
5. What are these invariant features in genres which display the purpose of the document?
6. How do human beings perceive genres/affordances of documents, what do they perceive and can their perceptions be modelled?
7. Can genre identification be performed by skimming methods affordance features?
8. How can genre and affordances be utilized in the relevance judgement of documents? What is the decision-making process?

4. PROPOSED EXPERIMENTS

We have demonstrated, in previous experiments [11, 39], that parsing shallow form features of XML technology in retrieval (text categorization) is effective and we now wish to ascertain whether the previous XML genre experimental features [8] will transfer to other technically or socially structured document collections, such as email and WWW media, in particular XHTML. A full description of our previous experiments and the classification accuracies can be found in my earlier work [40] The plan is to explore genre features which are maintained and exploitable in the explicit XML using other highly structured corpuses, such as Wikipedia.

We also intend to examine the possibility of creating or extending skimming models, such as those found in Mauldin's Ferret [34], Salton [41], DeJong [33] and Schank et al. [42].

Although affordances in ecological psychology, genre and IR are relatively popular and well-known areas of research, combining these areas constitutes a new approach. Our new approach will also include utilising eye-tracking experiments to obtain an accurate understanding of how humans view the invariant layout cues, such as white-space patterns or other formatting features which constitute genres and test whether genres act as 'affordances'. This eye-tracking technology, which can record eye-movements, should allow us to test our hypothesis that when we see a text, we perceive the invariants which indicate to us the purpose of the text. In this context, the genres act as affordances which allow a reader to make a decision regarding the relevance of the text and then add meaning later, which is a main tenet in Gibson's affordances, i.e. perception is followed by action This work has been inspired by the previous genre research carried out by Watt [18], and Campbell and Toms [27].

Two experiments have been planned so far: the first will test several genres of emails and the invariant features which are used to make decisions during the categorisation process. The second will be used on an arguably evolving set of genres which are inherent in the Wikipedia collection.

Finally, the structure which occurs naturally in social networks will be investigated, for example, in a community of practice, such as a Yahoo discussion room and Wikipedia. At present, the Yahoo forums offer very poor search facilities to allow the users to search through the archives of messages. One possible approach we intend to employ is that of testing retrieval on the Yahoo discussion forum corpus in which approximately 75000 messages have already been crawled. These messages are rich in natural language but the question is whether retrieval is improved by exploiting genre patterns and rules which normally emerge through social consensus in a community of practice and have a distinctive purpose and form.

5. CONCLUSIONS

We argue that the genre affords the purpose of the text to the reader: the invariant layout cues of the structured text serve as signposts to direct the reader towards the possible uses and purposes of the text. Categorising (or classifying) texts according to their genre features could result in significant improvements in text retrieval relevance and performance. For example, the abstract of an academic article allows the user to decide whether the article is actually useful and interesting and also allows the act of document filtration.

Schank et al. [42] in the field of artificial intelligence (A.I.) had a goal of making 'intelligent' machines, but pointed out that the creators of such systems disregard an area which should be at the forefront, that is, human psychology: a machine which is as intelligent as a human being must be able to emulate human behaviour. Our concept involves developing a machine which can perceive action and meaning, following Gibson's ideas of perception and affordances. Skimming retrieval is a perfect model to interlink these complex, but useful areas of research.

The BCS IRGS symposium offers an excellent opportunity for the examination and discussion of the potential uses of systems of this type.

REFERENCES

[1] Lalmas M, Ruthven I, A Model for Structured Document Retrieval: Empirical Investigations. *Hypermedia - Information Retrieval - Multimedia (HIM97)*. 1997:53-66.

[2] Wilkinson R. Effective retrieval of structured documents. In: Springer-Verlag New York INY, NY, USA, editor. SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval; 1994; Dublin 1994: Springer-Verlag New York, Inc. New York, NY, USA; 1994. p. 311-7.

[3] Dewdney N, VanEss-Dykema C, MacMillan R. The form is the substance: classification of genres in text. Association for Computational Linguistics Morristown, NJ, USA; 2001. p. 1-8.

[4] Allen RE, *Plato's Euthyphro and the Earlier Theory of Forms*. Prometheus Books 1970.

[5] Duff D, *Modern genre theory*. 1st ed: Longman; 2000.

[6] Bakhtin MM, *Speech Genres and Other Late Essays*. Austin,Texas: University of Texas Press; 1986.

[7] Yates J, Orlikowski WJ, Genres of organizational communication: a structurational approach to studying communication and media. *Academy of Management Review*. 1992;17(2):299-326.

[8] Clark MJ, Watt SNK. Classifying XML Documents by Using Genre Features TIR-07 4th International Workshop on Text-based Information Retrieval (DEXA 2007). Regensburg, Germany: IEEE; In Press.

[9] Gibson JJ, *The ecological approach to visual perception*. 2nd ed. New Jersey: LEA; 1986.

[10] Lalmas M, Rolleke T, Szlavik Z, Tombros T. Accessing XML documents: the INEX initiative. In: Agosti M, Fuhr N, editors. DELOS WP7 Workshop on the Evaluation of Digital Libraries; 2004 4-5 October 2004; University of Padua, Italy; 2004.

[11] Clark MJ. *Classifying XML Documents by Genre Vol. 1.* [MSc]. Aberdeen: The Robert Gordon University; 2005.

[12] Sebastiani F, Text Categorization. *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*. 2005:109--29.
[13] Rauber A, Muller-Kogler A. Integrating automatic genre analysis into digital libraries. JCDL '01: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries; 2001 June 24-28, 2001; Roanoke, Virginia: ACM Press; 2001.
[14] Meyer zu Eissen S, Stein B. Genre classification of web pages. Proceedings of the 27th German Conference on Artificial Intelligence (KI-2004), ; 2004; Ulm,Germany: Springer; 2004. p. 256-69.
[15] Andersen J. Genre - The Epistemological Lifeboat Epistemology and Philosophy of Science for Information Scientists. The Epistemological Lifeboat Epistemology and Philosophy of Science for Information Scientists 2005 13/06/06 [cited 2005; Available from: http://www.db.dk/jni/lifeboat/info.asp?subjectid=83
[16] Freedman A, Bazerman C, Medway P, *Genre and the new rhetoric*. Taylor & Francis, London,(1994); 1995.
[17] Breure L. Development of the Genre Concept. Information and Computing Sciences
University of Utrecht, The Netherlands; 2001.
[18] Watt S. Text Categorization Without Words: Can it Possibly Work. Aberdeen: The Robert Gordon University; 2004.
[19] Herring SC, Scheidt LA, Wright E, Bonus S, Weblogs as a bridging genre *Information, Technology & People*. 2005;18(2):142-71.
[20] Miller CR, Shepherd D, Blogging as Social Action: A Genre Analysis of the Weblog. *Into the Blogosphere: Rhetoric, Community, and the Culture of Weblogs*. 2004.
[21] Wenger E, *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press; 1999.
[22] Amini MR, Tombros A, Usunier N, Lalmas M, Gallinari P. Learning to summarise XML documents using content and structure. Proceedings of the 14th ACM international conference on Information and knowledge management; 2005 October 31–November 5, 2005; Bremen,Germany: ACM 2005. p. 297-8.
[23] Ramírez G, de Vries AP, Structural features in content oriented XML retrieval. *Proceedings of the 14th ACM international conference on Information and knowledge management*. 2005:291-2.
[24] Boese ES, Howe AE. Effects of web document evolution on genre classification. CIKM'05. Bremen: ACM Press New York, NY, USA; 2005. p. 632-9.
[25] Rehm G. Towards Automatic Web Genre Identification. Proceedings of the Hawai'i International Conference on System Sciences; 2002 January 7–10, 2002; Big Island, Hawaii: IEEE; 2002.
[26] Shepherd M, Watters C. Identifying Web Genre: Hitting A Moving Target. Proc of the WWW2004 Conference Workshop on Measuring Web Search Effectiveness: The User Perspective; 2004 18 May 2004.; New York; 2004.
[27] Campbell DG, Toms EG. Genre as interface metaphor: exploiting form and function in digital environments. Proceedings of the 32nd Hawaii International Conference on System Sciences 1999; Hawaii: IEEE; 1999.
[28] Finn A, Kushmerick N, Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*. 2003;57(11):1506-18.
[29] Shepherd M, Watters C. The Evolution of Cybergenres. In: Society IC, editor. HICSS '98: Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences; 1998; Hawaii: IEEE Computer Society; 1998. p. 97.
[30] Kennedy A, Shepherd M. Automatic Identification of Home Pages on the Web. 2005.
[31] Luštrek M. Overview of Automatic Genre Identification. Ljubljana, Slovenia: Jožef Stefan Institute, Department of Intelligent Systems; 2006 May 2006.
[32] Watt SNK. Text categorisation and genre in information retrieval. In: Goker A, Davies J, Graham M, editors. Information retrieval: Searching in the 21st Century: John Wiley & Sons; 2007.
[33] De Jong GF. *Skimming stories in real time: an experiment in integrated understanding.* [PhD]. New Haven: Yale University; 1979.
[34] Mauldin ML. Retrieval performance in Ferret a conceptual information retrieval system. 14th International Conference on Research and Development in Information Retrieval ACM SIGIR; 1991 October 1991; Chicago, U.S: ACM Press New York; 1991. p. 347-55.
[35] Watt S, Concept Linking for Information Integration in Open Book and Sentinel. 2003.
[36] Gellatly A, Braisby N, *Cognitive Psychology*. Milton Keynes: In association with The Open University; 2005.
[37] Zhang J. Categorization of Affordances. 2006 2006 [cited 2006 15 May]; Notes for published article]. Available from: http://acad88.sahs.uth.tmc.edu/courses/hi6301/affordance.html
[38] Zhang J, Patel VL, Distributed cognition, representation, and affordance. *Special issue of Pragmatics & Cognition*. 2006;14(2):333–41.
[39] Clark MJ. *Classifying XML Documents by Genre Vol. 2.* [MSc]. Aberdeen: The Robert Gordon University; 2005.
[40] Clark M. Structured text retrieval by means of affordances and genre. BCS IRSG Symposium: Future Directions in Information Access 2007. Glasgow, Scotland: BCS; 2007.
[41] Salton G, *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA 1971.
[42] Schank RC, Riesbeck CK, *Inside Computer Understanding: five programs plus miniatures*. 1st ed. Mahwah, NJ, USA Lawrence Erlbaum Associates; 1981.