Classifying Document Titles Based on Information Inference

Dawei Song[1], Peter Bruza[1], Zi Huang[2], Raymond Lau[3]

[1] Distributed Systems Technology Centre
The University of Queensland, QLD 4072 Australia
{dsong, bruza}@dstc.edu.au
[2] School of Information Technologies and Electronic Engineering
The University of Queensland, QLD 4072 Australia
huang@itee.uq.edu.au
[3] Centre for Information Technology Innovation
Faculty of Information Technology
Queensland University of Technology
GPO Box 2434, Brisbane, Q4001, Australia
r.lau@qut.edu.au

Abstract. We propose an intelligent document title classification agent based on a theory of information inference. The information is represented as vectorial spaces computed by a cognitively motivated model, namely Hyperspace Analogue to Language (HAL). A combination heuristic is used to combine a group of concepts into one single combination vector. Information inference can be performed on the HAL spaces via computing information flow between vectors or combination vectors. Based on this theory, a document title is treated as a combination vector by applying the combination heuristic to all the non-stop terms in the title. Two methodologies for learning and assigning categories to document titles are addressed. Experimental results on Reuters-21578 corpus show that our framework is promising and its performance achieves 71% of the upper bound (which is approximated by using whole documents).

1 Introduction

There has been ever expanding vast amount of electronic information available to us to process. In situations involving large amounts of incoming electronic information (e.g., defense intelligence), judgments about content (whether by automatic or manual means) are sometimes performed based simply on a *title* description or brief caption. For example, a human can quickly make the judgment that a web page title “Welcome to Penguin Books, U.K” refers to Penguin, the publisher. In regard to the following text "Linux Online: Why Linus chose a Penguin?", some agents can infer readily that "Linus" refers to Linus Torvalds, the inventor of Linux system, and the penguin mentioned here has to do with the Linux logo. In this way, the user can classify the incoming documents into different categories, e.g., “publisher”, “Linux”.

At first glance, the above examples show similarity to the Text Categorization (TC), whose task is to assign a number of pre-defined category labels to a document by learning a classification scheme from a set of labeled training data. The result of the learning is a classifier associated with each category C, which can decide whether an arbitrary document should be classified with C. A number of statistical learning algorithms have been investigated, such as K-nearest neighbour (KNN) [2], Naïve Bayes (NB) [7], Support Vector Machines (SVM) [11], and Neural Network (NN) [7], etc. These are supervised approaches- they make use of the manually pre-assigned categories associated with the training documents to train automatic classifiers for each predefined category. Many practical situations do not conform to the above format of text categorization. For example, labeled training may not be available, or the set of categories may not be predefined. As a consequence, unsupervised techniques are required in order to produce a classifier. In data intensive domains for example, defense intelligence, incoming documents may need to be classified on the basis of titles alone. Note that the categories like "publisher", "birds" and "Linux" may not explicitly appear in the document titles. In other words, they must be inferred.

In short, human agents can generally make robust judgments about what information fragments are, or are not about, even when the fragments are brief or incomplete. The process of making such "aboutness" judgments has been referred to as *informational inference* in our recent work [3, 9]. We have developed an information inference model which aims to mimick human judgments regarding a terse text fragment [3,9]. This model is a reflection of how strongly *Y* is *informationally contained* within *X*. The goal of this paper is to propose a framework to drive the use of information flow model for category learning and information classification on document titles.

2 Information Inference via Information Flow

2.1 Vector Representation of Information via HAL

A human encountering a new concept draws its meaning via an accumulation of experience of the contexts in which the concept appears. Burgess and Lund [6] developed a model called *Hyperspace Analogue to Language* (HAL), which automatically "learns" the meaning of a concept through how a concept appears within the context of other concepts. constructs a high dimensional semantic space from a corpus of text.

Given an *n*-word vocabulary, the HAL space is a $n \times n$ matrix constructed by moving a window of length *L* over the corpus by one word increment ignoring punctuation, sentence and paragraph boundaries. All words within the window are considered as co-occurring with each other with strengths inversely proportional to the distance between them. After traversing the corpus, an accumulated co-occurrence matrix for all the words in a target vocabulary is produced. The row and column vectors of every word record the co-occurrence information for other words preceding and following it. In our work, the row and column vectors in the HAL matrix corresponding to a word are added to produce a single vector representation for that word. Quality properties

are identified as those dimensions in the HAL vector for which are above a certain threshold (e.g., above the average weight within that vector). HAL vectors are normalized to unit length. For example, Table 1 lists part of the normalized HAL vector for "*Iran*" computed derived from applying the HAL method to the Reuters-21578 collection with stop words removed. The weights represent the strengths of association between "iran" and other words seen in the context of the sliding window: the higher the weight of a word, the more it has lexically co-occurred with "iran" in the same context(s). The dimensions reflect aspects which were relevant to the respective concepts during the mid to late eighties. For example, Iran was involved in a war with Iraq, and president Reagan was involved in an arms scandal involving Iran.

Table 1. HAL vector.

Iran	
Dimension	Value
arms	0.64
iraq	0.28
scandal	0.22
gulf	0.18
war	0.18
sales	0.18
attack	0.17
oil	0.16
offensive	0.12
missiles	0.10
reagan	0.09
...	...

The quality of HAL vectors is influenced by the window size; the longer the window, the higher the chance of representing spurious associations between terms. A window size of eight or ten has been used in various studies [3, 6, 9, 10]. We choose eight for the experiments reported in this paper.

More formally, a concept c is a vector representation: $c = < w_{cp_1}, w_{cp_2}, ..., w_{cp_n} >$ where $p_1, p_2, ..p_n$ are called dimensions of c, n is the dimensionality of the HAL space, and w_{cp_i} denotes the weight of p_i in the vector representation of c. A dimension is termed a property if its weight is greater than zero. A property p_i of a concept c is a termed a *quality property* iff $w_{cp_i} > \partial$, where ∂ is a non-zero threshold value. Let $QP_\partial(c)$ denote the set of quality properties of concept c. $QP_\mu(c)$ will be used to denote the set of quality properties above mean value, and $QP(c)$ is short for $QP_0(c)$.

2.2 Computing Information Flow in HAL spaces

Barwise and Seligman [1] have proposed an account of information flow that provides a theoretical basis for establishing informational inferences between concepts. For example, *penguin, books |- publisher* denotes that the concept "publisher" is carried informationally by the combination of the concepts "penguin" and "books". Said otherwise, "publisher" *flows* informationally from "penguin" and "books". The degree of information flow is directly related to the degree of inclusion between the respective HAL vectors [10]. Inclusion is a relation $\lhd$ over HAL vectors. Total inclusion leads to maximum information flow. HAL-based information flow is defined as:

$$i_1, \ldots, i_k \mid\!- j \text{ iff } \text{degree}(\oplus c_i \lhd c_j) > \lambda$$

where c_i denotes the conceptual representation of token i, and λ is a threshold value. ($\oplus c_i$ refers to the combination of the HAL vectors $c_1, \ldots, c_k$ into a single vector representation representing the combined concept. Details of a concept combination heu-

ristic can be found be in [10]). The degree of inclusion is computed in terms of the ratio of intersecting quality properties of c_i and c_j to the number of quality properties in the source c_i:

$$\text{degree}(c_i \lhd c_j) = \frac{\sum_{p_l \in (QP_{\mu}(c_i) \wedge QP(c_j))} w_{c_i p_l}}{\sum_{p_k \in QP_{\mu}(c_i)} w_{c_i p_k}}$$

Table 2. Information flows from "Gatt talks".

Information Flows	Degree
gatt	1.00
trade	0.96
agreement	0.96
world	0.86
negotiations	0.85
talks	0.84
set	0.82
states	0.82
EC	0.81
japan	0.78
...	...

The underlying idea of this definition is to make sure that a majority of the most important quality properties of c_i appear in c_j. Note that information flow produces truly inferential character, i.e., concept *j* need not have a positive value in the vector c_i. Table 2 shows an example of information flow computation where the weights represent the degree of information flows derived from the combination of "GATT" **(**General Agreement on Tariffs & Trade, which is a forum for global trade talks) and "talks".

3 Classifying Document Titles via Information Flow Inference

The architecture of information flow based information classification agent is depicted in the following figure:

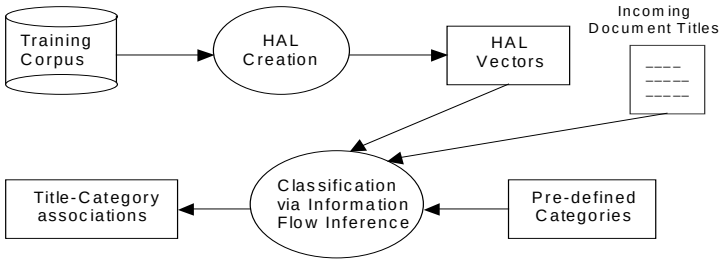

Fig. 2. Architecture of the intelligent classification agent

3.1 HAL Construction

A weighted vector represented HAL space will be produced from a large scale training collection. Once the HAL space has been created, it is ready for use by the information inference module. The training corpus may be dynamic – it could be expanded

when new information comes in. The changes of HAL vectors from the new data can then be updated accordingly. This module can be pre-processed and kept updated in the background.

3.2 Classification via Information Flow Inference

The HAL space will then feature a module of informational inference, which is sensitive to the context of local collection, which refers to incoming document titles. The information flow theory introduced in Section 2 allows information classification to be considered in terms of information inference.

3.2.1 Methodology-1

Suppose the user has a list of pre-defined categories. Consider the terms $i_{1,} \ldots, i_n$ being drawn from the title of an incoming document *D* to be classified. The concept combination heuristic can be used to obtain a combination vector $i_1 \oplus i_2 \oplus \ldots \oplus i_n$. The degree of information flow from $i_1 \oplus i_2 \oplus \ldots \oplus i_n$ to a category *j* can be calculated. If it is sufficiently high, category *j* can be assigned to *D*.

For example, "trade" is inferred from "GATT TALKS" with a high degree 0.96. The document titled ""GATT TALKS" can then be classified with category "trade".

3.2.2 Methodology-2

A variation of Methodology-1 can also be applied. A set of information flows can be computed from the document title terms $i_{1,} \ldots, i_n$. Similarly, a set of information flows can be computed on the basis of the category *j*. This is indeed a category learning process. The respective sets of information flows can then be matched. If there is sufficient overlap between the respective sets of information flows, category *j* can be assigned to document *D*. For example, the information flows from category "trade" is:

trade |- < trade:1.000 U.S.:0.967 japan:0.878 market:0.871 foreign:0.865 government:0.843 countries:0.837 world:0.816 economic:0.810 officials:0.810 dlrs:0.801 … >

By comparing the information flows from the document title "GATT TALKS" (see section 2) and the category "trade", we can find that they have a large overlap, for example, "trade", "japan", "world", "states", etc. Such overlap provides the basis of assigning the category "trade" to the document in question, even though "trade" does not explicitly appears in the document title.

4. Experiments

The effectiveness of this module is evaluated empirically on the Reuters-21578 corpus, which has been a commonly used benchmark for text classification. This cor-

pus consists of 21578 Reuters news articles, among which 14688 are training documents and the rest are test documents. The human indexers have pre-defined 120 topics (categories) and labeled the documents using the corresponding topics assigned to them.

Table 3. Test topics.

Topics	Number of relevant documents
acq (acquisition)	719
coconut	2
coffee	28
crude	189
grain	149
interest	133
nickel	1
oat	6
peseta	0
plywood	0
rice	24
rupiah	0
rye	1
ship	89
sugar	36
tea	4
trade	118

In our experiments, seventeen categories (topics) were selected from the total of 120 topics. The 14,688 training documents were extracted, and pre-labeled topics were removed, i.e., there won't be any explicit topic information in the training set. This training set is then used to construct a HAL space from which information flows can be computed. After removing stop words, the total size of vocabulary for the training set is 35,860 terms.

In addition, a collection of 1,394 test documents, each of which contains at least one of the 17 selected topics, is formed. Similarly, the topic labels are removed. With respect the test set, only 14 topics have relevance information (i.e., assigned to at least one test document). Among the 14 topics, five have above 100 relevant documents and four have below 10 relevant documents. The average number of relevant documents over the 14 topics is 107. We chose this set of topics because they vary from the most frequently used topics in Reuters collection like "acquisition" to some rarely used ones such as "rye". Note that we use the real English word "acquisition" instead of the original topic "acq" for information flow computation. Table 3 lists the selected topics and their relevance information.

Experiment 1 – Effectiveness of Methodology 1

The aim is this experiment is to test the effectiveness of deriving categories directly from document titles information flow model. We use only titles of test documents. The average title length is 5.38 words. The concept combination is applied to each title to build title vector. The top 80 Information flows with associated degrees are derived from each title vector. If a topic appears in the list of information flows of a title, it is then considered to be relevant to this title. Our previous experiments in query expansion via information flow shows that keeping top 80 information flows produces the best results [3]. This parameter setting will be used throughout this paper. Figure 3 illustrates this experiment.

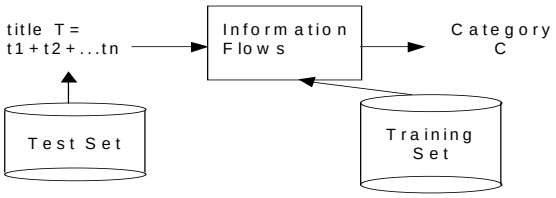

Fig. 3. Methodology of Experiment 1

Experiment 2 - Effectiveness of Methodology 2

The aim of this experiment is to use information flow model to expand document titles and categories, instead of deriving categories from titles directly as in experiment 1. Similarly only titles of test documents are used. The top 80 Information flows are derived from each title vector, and then used to match top 80 information flows of each topic using dot product function. Figure 4 summaries experiment 2.

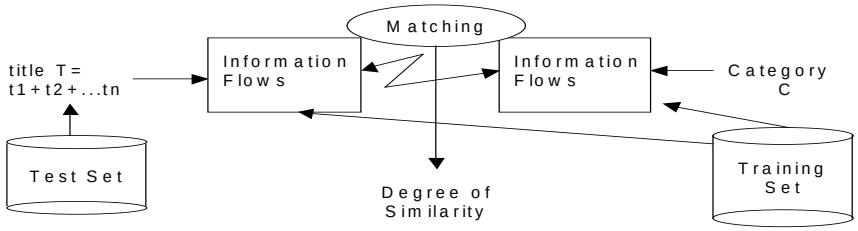

Fig. 4. Methodology of Experiment 2.

Experiment 3 – Upper bound

The upper bound of document title classification module can be roughly measured by using the whole document to match the categories. In our case, this involves computing the information flows on the basis of a category *j*, and using the resultant information flows as expansion terms. These can then be matched with the document *D*. If the match is deemed sufficient, *D* can be classified with *j*.

The test documents are indexed using the document term frequency and inverse collection frequency components of the Okapi BM-25 (Robertson et al. 1995) formula. The average document length in the test set is 77.7 words. They are matched against the top 80 information flows of each topic using dot product. The methodology of experiment 3 is visualized as below:

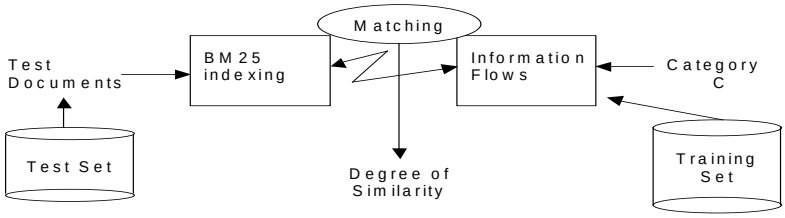

Fig. 5. Methodology of Experiment 3.

Experimental results

The performance measures used in the paper include interpolated 11-point Average Precision[1], Initial precision (i.e. interpolated precision at 0% recall) and R-precision (i.e. precision after R (= number of documents which actually belong to a category) documents have been retrieved). We choose them because we think people care more about the precision of the top ranked documents with respect to a category. In our experiments, the documents are ranked by their similarity to the topics. The experiment results are shown in Table 4 and Figure 6.

Table 3. Comparison of performance between three experiments.

	Avg Precision	Initial Precision	*R-Precision*
Experiment-1	0.461	0.786	0.454
Experiment-2	0.404	0.819	0.406
Experiment-3	0.721	0.911	0.640

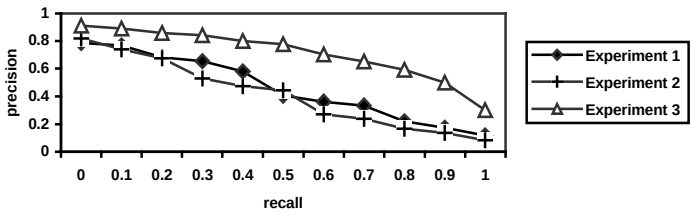

Fig. 6. Precision-Recall curves of three experiments.

Discussion

Experiment 3 serves as a baseline and sets a somehow upper bound for the other two experiments. Compared to the baseline, Experiment 1 and 2 both performs reasonably well. In particular, the precision values at low recall points 0.0, 0.1 and 0.3 are pretty high. Their R-precisions achieved separately 71% and 64% of the upper bound. This indicates the information flow model seems promising for reasoning about terse text fragments.

Experiment 1, which applies information flow model to derive categories from titles directly, performs better than experiment 2, which is based on the similarity between expanded titles and categories. This indicates that the document title classification could be considered more like an inference process rather than similarity.

---

[1] Precision is the proportion of documents classified with a category that really belong to that category. Recall is the proportion of the documents belonging to a category that are actually assigned that category. The average precision is computed across at 11 evenly spaced recall points (0, 0.1, 0.2, …1.0), and averaged over all the queries (categories). This corresponds to micro averaged precision in the text categorization field.

5. Conclusions and future Work

We have proposed an intelligent agent for classifying incoming document titles, based on a theory of information inference. The information is represented as vector spaces computed by the Hyperspace Analogue to Language (HAL) model. Information inference can be performed on the HAL spaces via computing information flow between vectors or combination vectors. Based on this theory, a document title is treated as a combination vector by combining all the non-stop title terms. A number of pre-defined categories, which are among the set of top-ranked information flows inferred from the combination vector, can be assigned to this title. Alternatively, the categories can also be represented as HAL vectors. Information flows inferred from category vectors are used to match against the information flows inferred from document title vectors. The categories with sufficiently high similarity scores are assigned to the corresponding document titles. Experimental results show that the performance achieves 71% of the upper bound (which is approximated by using whole documents).

Acknowledgements

The work reported in this paper has been funded in part by the Cooperative Research Centres Program through the Department of the Prime Minister and Cabinet of Australia.

References

1. Barwise, J. and Seligman, J. (1997) *Information Flow: The Logic of Distributed Systems.* Cambridge Tracts in Theoretical Computer Science, 44.
2. Dasarathy, B.V. (1991) *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* IEEE Computer Society Press.
3. Bruza, P.D. and Song, D. (2002) Inferring Query Models by Computing Information Flow. *Proceedings of the 12th International Conference on Information and knowledge Management (CIKM2002)*, pp. 260-269.
4. Burgess, C., Livesay, K. and Lund K. (1998) Explorations in Context Space: Words, Sentences, Discourse. *Discourse Processes, 25(2&3), 211-257.*
5. Gärdenfors, P. (2000) *Conceptual Spaces: The Geometry of Thought.* MIT Press.
6. Lund, K. and Burgess C. (1996) Producing High-dimensional Semantic Spaces from Lexical Co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28(2), 203-208.*
7. Mitchell, T. (1996) *Machine Learning.* McGraw Hill.
8. Robertson, S.E., Walker, S., Spark-Jones, K., Hancock-Beaulieu, M.M., and Gatford, M. (1995) OKAPI at TREC-3. In *Proceedings of the 3rd Text Retrieval Conference (TREC-3).*
9. Song, D. and Bruza, P.D. (2001) Discovering Information Flow Using a High Dimensional Conceptual Space. In Proceedings of the *24th Annual International Conference on Research and Development in Information Retrieval (SIGIR'01*), pp. 327-333.
10. Song, D. and Bruza, P.D. (2003) Towards A Theory of Context Sensitive Informational Inference. *Journal of American Society for Information Science and Technology, 54(4). pp. 326-339.*
11. Vapnic, V. (1995) The Nature of Statistical Learning Theory. Springer.