The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1468-4527.htm

A ranking algorithm for query expansion based on the term's appearing probability in the single document

Shihchieh Chou and Chinyi Cheng
Department of Information Management, National Central University, Chung-Li, Taiwan, Republic of China, and
Szujui Huang
Inventec Corporation, Chung-Li, Taiwan, Republic of China

Refereed article received
5 August 2009
Approved for publication
13 July 2010

Abstract

Purpose – The purpose of this paper is to establish a new approach for solving the expansion term problem.

Design/methodology/approach – This study develops an expansion term weighting function derived from the valuable concepts used by previous approaches. These concepts include probability measurement, adjustment according to situations, and summation of weights. Formal tests have been conducted to compare the proposed weighting function with the baseline ranking model and other weighting functions.

Findings – The results reveal stable performance by the proposed expansion term weighting function. It proves more effective than the baseline ranking model and outperforms other weighting functions.

Research limitations/implications – The paper finds that testing additional data sets and potential applications to real working situations is required before the generalisability and superiority of the proposed expansion term weighting function can be asserted.

Originality/value – Stable performance and an acceptable level of effectiveness for the proposed expansion term weighting function indicate the potential for further study and development of this approach. This would add to the current methods studied by the information retrieval community for culling information from documents.

Keywords Query languages, Information retrieval, Information management

Paper type Research paper

Introduction

Easily accessed publishing channels have resulted in a serious problem: information overload. Research suggests that a traditional keyword-based query is inefficient (Furnas *et al.*, 1988; Carpineto *et al.*, 2002; Hariri, 2008). One source of this problem is word mismatch.

In a keyword-based query, word mismatch denotes a situation wherein the words people use to describe the concepts found in queries are different from those used by

This work was supported in part by the National Science Council, Taiwan, under Grant No. NSC98-2410-H-008-028-MY2.

Online Information Review
Vol. 35 No. 2, 2011
pp. 217-236
© Emerald Group Publishing Limited
1468-4527
DOI 10.1108/14684521111128014

authors to describe the same concepts in their documents (Furnas *et al.*, 1988; Xu and Croft, 2000). This situation becomes even more problematic when queries are short (Carpineto *et al.*, 2001). Unfortunately people most often use three or fewer keywords to formulate their queries (Chau *et al.*, 2005).

Query expansion that adds adequate keywords to the original query is one way to address the problem. For example a user who makes an initial query with the keywords tropical storms might be interested in the damage caused by tropical storms. The retrieved documents may also discuss the natural phenomena of tropical storms. Let us assume the user selects the retrieved documents that describe the damage caused by tropical storms as relevant and the term damage appears with high frequency in the relevant documents. To enhance retrieval effectiveness it is reasonable to add the term damage to the initial query for further searching.

Studies of query expansion have usually employed a weighting function that utilises the information residing in the retrieved relevant/non-relevant documents to re-weight the candidate expansion terms for keyword selection. Most of these studies can be classified as one of two important approaches: the vector space model and the probabilistic model (Harman, 1992; Carpineto *et al.*, 2001; Alshaar, 2008).

This paper is concerned with the development of a weighting function using a new approach. It was initially inspired by Harter's concept of the specialty word. In an earlier study Harter (1975a, b) determined that a specialty word tended to appear more densely or with a higher probability in a few elite documents. Based on this observation it is natural to select a term for expansion by checking the term's appearance probability. However, before the final selection, there are questions that need to be dealt with:

(1) How do we select terms with the same appearance probability, but appearing in documents differently ranked by relevance?

(2) How do we select terms with the same appearance probability, but with different appearance situations in terms of the number of documents in which the term has appeared?

Generally speaking, the question asked is this: how do we address a situation in which terms with the same appearance probability are used but in reality do not have equal importance?

In the past the two important approaches did not directly deal with these matters despite the fact that they were based on an assumption similar to the one proposed by Harter. The vector space model approach dealt with term frequency. The primary concepts included the adjustment of the frequency according to situations and the summation of weights. The probabilistic model approach dealt with term appearance probability. The central concepts included the measurement of probability and the comparison of probabilities.

It is not necessary to address these matters directly in order to mitigate the expansion term problem, however direct manipulation could prove a worthwhile approach for solving this problem. Because the concepts behind the two successfully applied approaches have proven valuable, we are concerned with the combined application of both in regard to the direct manipulation of the aforementioned matters. By employing the concept of probability measurement as well as applying the concepts of adjustment according to situations and summation of weights, it is possible to form an appropriate logic that can deal with the abovementioned directly. If a new approach could be successfully established, the solution to the expansion term problem would be expanded. Moreover the justified logic would benefit the study of information retrieval.

Hence this study aims to establish a new approach for solving the expansion term problem by combining some of the valuable concepts used in previous approaches. More specifically this study seeks to develop an expansion term weighting function by the combined application of the concepts of probability measurement, adjustment according to situations, and summation of weights. The proposed approach maintains the probability measurement as the base but admits the adjustment and summation of the probability values. The final output can be termed pseudo probability.

The remainder of this paper is organized as follows. In the following section we introduce the background and the framework of this study. Then we review the contextual literature for the two main approaches to the expansion term problem in order to conceptualise our application and compare previous work with our proposition. In the next section we describe the development of the proposed expansion term weighting function. In the subsequent section we evaluate the performance of the proposed expansion term weighting function. We first compare it with the baseline ranking model and then compare it with other weighting functions. Finally we draw conclusions and offer recommendations for future research.

Local analysis-based query expansion

Query expansion is a technique that helps users reformulate their queries with additional terms that are related to their original query to enhance the effectiveness of information retrieval; the process can be interactive or automatic. In interactive mode the user is required to identify the retrieved documents as relevant or non-relevant (Rocchio, 1971). By employing a term weighting function, candidate expansion terms are re-weighted and selected from the relevant documents. In automatic mode users are not asked to make any effort in the process of query expansion (Xu and Croft, 2000). This can be divided into global analysis and local analysis according to the source.

Global analysis based query expansion selects expansion terms by utilising the term relationships that are extracted from the whole document collection (Spärck Jones, 1971; Deerwester *et al.*, 1990; Qiu and Frei, 1993; Jing and Croft, 1994; Schütze and Pedersen, 1994) or the user's log (Cui *et al.*, 2003). Term clustering, the earliest global analysis technique, examines the co-occurrence of terms and groups the terms with high similarity into the same cluster (Spärck Jones, 1971). Latent semantic indexing, originally used for the dimension reduction of a term vector (Deerwester *et al.*, 1990), is aimed at discovering the correlations among terms. Formal concept analysis (Carpineto and Romano, 2000), Phrasefinder (Jing and Croft, 1994), and similarity thesauri (Qiu and Frei, 1993) analyse and apply synonyms, hypernyms and hyponyms of terms. The whole document set must be examined and analysed in order to use global analysis techniques.

Unlike global analysis, local analysis-based query expansion selects expansion terms from a number of top-returned documents that are set as pseudo-relevant. The idea of using a set of retrieved documents for the selection of expansion terms was proposed in the 1970s by Attar and Fraenkel (1977) as well as Croft and Harper (1979). This approach is thought to be simple and is more efficient in large document collections. Some research has indicated that local analysis is competitive with global analysis (Efthimiadis and Biron, 1994; Evans and Lefferts, 1994; Buckley *et al.*, 1994; Robertson *et al.*, 1995; Xu and Croft, 1996). Recently terms such as local feedback or pseudo-relevance feedback also refer to this application (Chirita *et al.*, 2007).

In local analysis-based query expansion, expanding term re-weighting stands at the core of the effort. Most research has applied Rocchio's (1971) re-weighting framework to work on it (Srinivasan, 1996; Mitra *et al.*, 1998; Singhal *et al.*, 1999). Many empirical studies have developed and compared the effectiveness of different term weighting functions utilised in the re-weighting framework (Efthimiadis, 1993, 1995; Carpineto *et al.*, 2001, 2002). These related studies demonstrate the importance of the term weighting function within local analysis based query expansion.

In this study we focus on local analysis based query expansion. The typical procedure is as follows:

1. The retrieval system executes the first-pass retrieval and ranks documents from the document base through an initial query. In this step a document-query weighting model is used to first compare the similarities between documents and the query and then the retrieval system will rank the retrieved documents.
2. A number (R) of retrieved and ranked documents are set as pseudo-relevant documents.
3. The retrieval system selects a number (E) of expansion terms from the pseudo-relevant documents and adds them to the initial query to form the expansion query. In this step the retrieval system re-weights the candidate expansion terms within a re-weighting framework, such as Rocchio's (1971) or its variant version. In the re-weighting framework a weighting function will be applied to modify the weights of the candidate expansion terms. After re-weighting of the candidate expansion terms the retrieval system will rank and select these terms.
4. The retrieval system uses the newly selected terms in the expanded query to execute the second-pass retrieval to retrieve and rank documents.

In the above procedure the following affects the performance of the retrieval: the number of pseudo-relevant documents (R), the number of expansion terms (E), the re-weighting framework, the constants in the re-weighting framework, and the weighting function in the re-weighting framework. In this study we tested the R value, the E value and the constants in the re-weighting framework and set them as system variables. We used the reduced version of Rocchio's (1971) re-weighting framework proposed by Carpineto *et al.* (2002) as shown in equation 1. We also developed the weighting function for testing:

$$W_{t,Q\exp} = \alpha \cdot W_{t,Qun\exp} + \beta \cdot score_t \tag{1}$$

where:

$W_{t,Q\exp}$: the weight assigned to term t after query expansion.

$W_{t,Qun\exp}$: the weight of term t in the unexpanded query.

$score_t$: the value assigned to term t by the term weighting function being used.

α and β: constants to control the relative contributions of the unexpanded query terms and the candidate expansion terms.

As mentioned earlier, whether interactive or automatic, the term weighting function plays an important role in the re-weighting of the candidate expansion term. The primary purpose of the term weighting function is to provide the candidate expansion term with an appropriate value so as to assist the selection of an adequate expansion term. Owing to the potential of the term weighting function in the enhancement of information retrieval, much study has been devoted to its development. These studies can commonly be classified as taking either the vector space model approach or the probabilistic model approach (Harman, 1992; Carpineto *et al.*, 2001; Alshaar, 2008).

The vector space model approach deals mainly with frequency (Harman, 1992; Alshaar, 2008). The most well-known study using this approach was conducted by Rocchio (1971). Rocchio's solution was to adjust a query vector by the term appearance frequency in the relevant/non-relevant documents retrieved. The primary operation was to have the terms of the query vector re-weighted by adding the weights of those terms occurring in the relevant documents and subtracting the weights of those terms occurring in the non-relevant documents.

Rocchio's study established the valuable concepts of adjustment and summation. The former refers to the change of the term's weight according to situations. In Rocchio's solution a term's appearance is examined in relevant or non-relevant documents. The weight of a term is adjusted according to the term's appearance situation. The latter refers to the total power of a term as it competes as an expansion term. Rocchio's study shows that the arithmetic sum of a term's weight can properly reflect its power as an expansion term.

Most studies using the vector space model approach applied the useful concepts as mentioned while implementing varied strategies for adjustment. The following typify some of the varied strategies:

- Modify the constant values of Rocchio's formula. Studies conducted by Yu *et al.* (1976), Koster and Beney (2007), and Moschitti (2003) changed the values for the adjustment constants in order to achieve better retrieval accuracy.
- Apply the document ranking information for adjustment. Rocchio's (1971) formula did not consider the application of the document ranking information derived from relevance ratings. Studies by Balabanovic (1997) and Nick and Themis (2001) collected the information pertaining to document ranking and applied it to adjust the weight of the query term.
- Calculate the relevance degree for adjustment. Kim *et al.* (2001) calculated the terms' relevance degree using term co-occurrence similarity and fuzzy inference. Then they used the calculated relevance degrees to adjust the terms' weights by adding their relevance degrees to their initial weights.
- Make use of experts' relevance feedback for adjustment. Azimi-Sadjadi *et al.* (2004) developed a learning mechanism to optimally map the original query using relevance feedback from multiple expert users.
- Refer to the user's preference for adjustment. Shanfeng *et al.* (2001) combined user preference with relevance feedback to rank web pages.

In our review of these studies we have distilled the concept of adjustment and shown how it has been applied. In light of these salient works it becomes more feasible for us to adopt the two concepts of adjustment and summation for the development of the expansion term weighting function.

The probabilistic model approach deals mainly with term appearance probability. Harter's (1975a, b) 2-Poisson model was one of the very first to use this approach (Amati and van Rijsbergen, 2002). Harter observed that the distribution of specialty words over a text collection deviated from the distributional behaviour of non-specialty words. While non-specialty words were modelled by a Poisson distribution with mean λ, a specialty word could be mechanically detected by measuring the extent to which its distribution deviated from a Poisson distribution. Harter used the Poisson distribution only to select high-quality indexing words. Robertson and Spärck Jones (1976) explored the potential effectiveness of Harter's model for direct retrieval exploitation. Developed in 1976 their first probabilistic model, known as the binary independence retrieval model, estimated the probability that each index term would be represented in a relevant document. To do this each index term was given a weight based on its presence and absence probabilities in relevant and non-relevant documents. Doszkocs (1978) (Carpineto *et al.*, 2001) produced another early work in which the appearance probabilities of the candidate expansion terms in relevant documents and the whole document set are analysed to suggest relevant terms for query expansion.

Over a decade later Harman (1992) revisited several important studies that included Robertson and Spärck Jones (1976), Doszkocs (1978), Croft and Harper (1979), and Porter and Galpin (1988). Harman's report presented over a decade's worth of major studies of the probabilistic approach. Porter and Galpin's (1988) simple formula provided a snapshot of these works. The main idea of the formula was to calculate the competing score for the candidate expansion term by subtracting the term appearance ratio in the whole set of documents from the term appearance ratio in the relevant documents.

Drawing from Harter's (1975a, b) early works on Poisson distribution, Amati and van Rijsbergen (2002) created a framework for deriving term-weighting models by measuring the divergence of the actual term distribution from that obtained through a random process. This study reflected another decade's research and study concerns. The Carpineto *et al.* (2001) weighting formula, which will be compared to our proposed weighting function later in this paper, is an example of the studies from that period. The expansion term weighting function of Carpineto *et al.*'s proposition, as shown in equation 2, measured the probabilities of candidate expansion terms in both the relevant and the whole document set by the Kullback-Leibler divergence (KLD). KLD, or the relative entropy, was defined in information theory (Losee, 1990; Cover and Thomas, 1991) and was used to measure the divergence or distance between two probability distributions:

$$weight_{KLD}(t) = [P_R(t)] \cdot \log \left[P_R(t)/P_C(t)\right] \tag{2}$$

where:

$weight_{KLD}(t)$: the weight for the candidate expansion term t.

$P_R(t)$: the probability of the candidate expansion term t in relevant documents.

$P_C(t)$: the probability of the candidate expansion term t in whole documents.

By describing the studies of the probabilistic approach, we have presented a general view of how term appearance probability was dealt with. The focus of the probabilistic approach has been described as the estimation of the probability that a document was relevant to the user (Alshaar, 2008).

In comparing the two approaches to expansion term weighting, the fundamental assumptions of term appearance deviation are similar, although the management of each is totally different. The major concepts of the probability approach include probability measurement and probability comparison in distribution, distinction, divergence, or distance. It focuses on a comparison of probability deviation for term weighting. The underlying information for terms with the same probability but in different appearance situations are not managed directly. In contrast the frequency approach focuses on situation management for term weighting.

Proposition for an expansion term weighting function

This study proposes an approach that begins with the concept of probability measurement and applies the concepts of adjustment according to situations and summation of weights in order to develop the expansion term weighting function. This approach originates in Harter's specialty word concept. According to Harter (1975a, b) specialty words, being informative, tend to appear more densely in a few elite documents, whereas non-specialty words are more likely to be randomly distributed over the collection. Adopting Harter's proposition we assume that a term which appears more frequently or densely in a single document contained in the pseudo-relevant document set could be informative. In other words this term may be adequate for an expanding query. Considering the studies reviewed earlier we propose that the measurement of the appearance probability of a term in a single document lays a solid foundation for the selection of the expansion term. From this base we then exploit the concept of adjustment to manage a term's appearance probability according to situations and the concept of summation to amount a term's final power in the competing as an expansion term. According to the proposition our development of the expansion term weighting function has gone through a specific series of considerations as follows.

First is the weighting for the terms with different appearance probabilities in the single document belonging to the pseudo-relevant document set. Applying the basic assumption of the vector space model for dealing with frequency, we set a term with a higher appearance probability to a higher weight. In the weighting function this weighting rule is determined by using the term's appearance probability: $P_{di}(t_i)$, where $P_{di}(t_i)$ is the appearance probability of term t_i in the pseudo-relevant document d_i; d_i is the document i belonging to the pseudo-relevant document set R; and t_i is the term i.

Second is the weighting for the terms that appear in the pseudo-relevant documents differently in terms of the number of the documents in which a term has appeared. Adopting the concept of summation as mentioned earlier, we sum a term's appearance probabilities in separate pseudo-relevant documents to its weight. Using this calculation we give a higher weight to the terms that appear in more pseudo-relevant documents. In the weighting function this weighting rule is accomplished by summing the term's appearance probabilities as: $\sum_{d_i \in R} P_{d_i}(t_i)$, where $\sum_{d_i} \in R$ sums term i's appearance probabilities in documents d_i that belong to the pseudo-relevant document set R.

Third is the weighting for the terms as the two previous weighting situations are considered in combination. When we have combined the above two weighting situations a term can be classified into one of the following four situations:

(1) high appearance probability in the single document and appears in more documents;

(2) high appearance probability in the single document and appears in fewer documents;

(3) low appearance probability in the single document and appears in more documents; and

(4) low appearance probability in the single document and appears in fewer documents.

In reality a term will stand somewhere along lines of continuities in relation to the four situations. The two weighting rules set earlier serve as the two axes in a determination of the term's weight.

The question of concern is: in the determination of a term's ranking priority for query expansion, should the two axes' powers be treated as equal? Consider terms A and B that have equal appearance probabilities in the pseudo-relevant document set. Term A appears with a high probability in one document only, and term B appears not only once in a single document but in many documents. Recalling the specialty word concept and the studies reviewed earlier, term A would undoubtedly be more informative than term B as the expansion term. This extreme case shows that the axis of the term's appearance probability in a single document will receive more emphasis than the other. In other words the accumulated weight for a term that appears in more pseudo-relevant documents must be reduced to decrease its effect on the determination of the expansion term.

Based on an analysis of the above we lower the weight of the term as the term appears in more pseudo-relevant documents in order to decrease the accumulated weight of the term. In the weighting function this weighting rule is accomplished by having the term's appearance probability $P_{di}(t_i)$ time: $\log_2(P_{d_i}(t_i)/P_R(t_i))$, where $P_{di}(t_i)$ is the appearance probability of term t_i in the pseudo-relevant document d_i, and $P_R(t_i)$ is the appearance probability of term t_i in the whole pseudo-relevant document set. To the quantum of: $P_{d_i}(t_i)/P_R(t_i)$, we finally take the log transformation to narrow the difference.

Fourth is the weighting for the terms appearing in documents that have different comparison values of document-query similarity. We propose the application of a similarity comparison between document and query for weighting for two reasons. The first is that in information retrieval, similarity comparisons are usually calculated between document and document, document and user profile, or document and query, and the weighting of terms based on document ranking by similarity comparison has been effectively applied in many studies (Balabanovic, 1997; Nick and Themis, 2001). The second reason is that in automatic query, the relevant documents are pseudo. That means there are non-relevant documents normally contained in the pseudo-relevant document set. Thus to distinguish the terms appearing in real relevant documents from those appearing in non-relevant documents in order to weight them differently is undoubtedly important. In automatic query the similarity comparison can be conducted using a document-query weighting model such as BM25 (Robertson *et al.*, 1998; Carpineto *et al.*, 2001). With this similarity calculation every document will stand somewhere along the continuum from relevant to non-relevant.

Accordingly, we adjust the weight of the term derived previously according to the document-query similarity comparison for the document in which the term appears. In the weighting function this weighting rule is accomplished by having the term's weight derived from the previous calculations time: $sim(d_i, q)/\sum_{d_i \in R} sim(d_i, q)$, where $sim(d_i,q)$ is the value of similarity comparison between document d_i and the initial query $q;\sum_{d_i \in R} sim(d_i, q)$ is the summation of all $sim(d_i,q)$ while d_i belongs to the pseudo-relevant document set R. The divisor of $\sum_{d_i \in R} sim(d_i, q)$ is applied here for normalisation.

Fifth is the weighting for the terms that appear in documents differently in terms of the number of the documents in which a term has appeared. The design of the weighting function so far takes into account the term's appearance situation in the pseudo-relevant document set only and does not consider the term's appearance in the whole document set. Numerous studies as reviewed earlier have effectively contrasted the term's appearance situation in the relevant documents to the whole document set. The goal of this is to reflect the term's divergence or specialty. In this study we use the ratio of the number of the term-appeared documents to the number of whole documents to represent the term's specialty. The argument is that as the term appears in more documents, it would have more opportunities to appear in non-relevant documents and there would be a greater possibility of the term being non-specialty.

With this understanding we decrease the weight of the term derived before if the term has appeared in more documents. In the weighting function this weighting rule is accomplished by having the term's weight derived from the previous calculations time: $\log_2(N/N_{t_i})/\log_2 N$, where N is the number of the documents in the whole collection; N_{t_i} is the number of the documents in which term t_i appears. The *log* is taken in the formula to narrow the difference and the log_2N is used for normalisation.

In summary, the weighting function of our design (which we call non-randomness-based weighting (NBW) reveals itself to be:

$$Weight(t_i) = \left(\sum_{d_i \in R} P_{d_i}(t_i) \times \log_2 \left(\frac{P_{d_i}(t_i)}{P_R(t_i)} \right) \times \frac{sim(d_i, q)}{\sum\limits_{d_i \in R} sim(d_i, q)} \right) \times \left(\frac{\log_2 (N/ > N_{t_i})}{\log_2 N} \right)$$

where:

$Weight(t_i)$: the weight assigned to term t_i.

R: the pseudo-relevant document set.

$P_{di}(t_i)$: the appearance probability of term t_i in the pseudo-relevant document d_i.

d_i: the document i belonging to the pseudo-relevant document set R.

t_i: the term i.

$P_R(t_i)$: the appearance probability of term t_i in the whole pseudo-relevant document set.

$sim(d_i,q)$: the value of the similarity comparison between document d_i and the initial query q.

N: the number of documents in the whole collection.

N_{t_i}: the number of documents in which term t_i appears.

Evaluation

To evaluate the performance of the proposed NBW expansion term weighting function, we conducted two experiments. In the first experiment the document retrieval effectiveness of the unexpanded query and the expanded query with NBW were compared. In the second experiment the document retrieval effectiveness for the expanded queries based on NBW and other expansion term weighting functions, including CHI-1, CHI-2 and KLD, were examined. Each application of the expansion term weighting functions was implemented within Rocchio's (1971) framework.

Data set and system implementation

We used a Cranfield collection as the test data set. The collection contained 1,400 documents on the subject of aerodynamics. It provided 225 test topics of documents with perceived relevance. Of the 225 queries 15 could not retrieve any document using NBW or any of the other three weighting functions. In the examination of the 15 queries we determined that there was only one document retrieved for each query after the first pass retrieval. Obviously these queries were not suitable for expansion and were excluded from the experiments. After refining the data set the removal of stop-words and stemming was completed using Terrier (Ounis *et al.*, 2005).

Terrier 1.0.2, developed at the University of Glasgow, was used as the platform for the development of our experimental retrieval system. It was executed in Java language with all functions encapsulated within Java classes. This allowed for easy development of the document retrieval system. In addition to supporting the development of the retrieval system, Terrier also supports the evaluation of retrieval performance. Once the query is made Terrier will engage in the matching process that applies the document-query weighting algorithm to score the document by its relevance to the query. Then Terrier will return the list of retrieved documents that are ordered by the scores of relevance.

Terrier has been used with many document ranking models, including BM25 which we used as a baseline. It also supported Rocchio's automatic query expansion framework. Since Terrier supported a TREC data set by default, we modified some of its settings to support the Cranfield collection. We also implemented the expansion term weighting functions that were not provided but were used for experiments.

Parameter study

With NBW applied to Rocchio's framework, four parameters could affect retrieval performance: the two constants of α and β in the framework and both the number of relevant documents (R value) and the number of expansion terms (E value). To maximise performance we conducted tests to determine the appropriate values for the parameters.

The α and β values in the framework were used to control the relative contributions of the original query and the pseudo-relevant document set in the weighting of the expansion terms. If the pseudo-relevant documents were well qualified in terms of real relevance, the terms belonging to that set should be more likely to be selected for expansion than the terms of the original query. In this case the pseudo-relevant document set should receive more emphasis in contrast to the original query during the weighting of the expansion term by giving a higher value to β (Buckley *et al.*, 1994; Carpineto *et al.*, 2002). In other words the two values of α and β reflected the relative contributions of the original query and the pseudo-relevant document set. Therefore either of the two values could be fixed and another value adjusted to detect an adequate ratio. Since this study focuses on the effect of the expansion term, we chose to set the value of α at 1 to keep the original query typical and the value of β adjusted from 1.1 to 1.9 (step 0.2), while having the other two parameters of R and E controlled by setting them at 10 and 40, respectively, after various pretests.

The two standard TREC evaluation measures – average precision (AV-PREC) and R-precision (R-PREC) – were used in the tests (Manning *et al.*, 2008). Table I presents the retrieval performance for average precision and R-precision as regards the value of β changes. The results show that NBW performs best when β is set to 1.5.

To detect the best values for R and E during maximisation of the retrieval performance, we tested various combinations of the two values while setting the value of α to 1 and the value of β to 1.5. The value of R was tested from 5 to 20 (step 5) while the value of E was changed from 10 to 50 (step 10). The average precision was used to measure retrieval performance. Table II shows that the combination of setting the R value to 10 and the E value to 40 provided the best performance, achieving 0.4608 in average precision.

Based on the above tests the values of the four parameters were set to the following for the two experiments: $\alpha = 1$, $\beta = 1.5$, $R = 10$, and $E = 40$.

β	AV-PREC	R-PREC
1.1	0.4586	0.4357
1.3	0.4594	0.4355
1.5	0.4608	0.4379
1.7	0.4604	0.4353
1.9	0.4598	0.4367

Notes: $\alpha = 1, R = 10, E = 40$

Table I. AV-PREC and R-PREC for different values of β

	E-10	E-20	E-30	E-40	E-50
R-5	0.4249	0.4526	0.4496	0.4528	0.4505
R-10	0.4300	0.4558	0.4546	0.4608	0.4566
R-15	0.4254	0.4506	0.4500	0.4508	0.4522
R-20	0.4218	0.4439	0.4459	0.4483	0.4466

Table II. AV-PREC for different combinations of R and E values

Experiment 1

In order to produce an overview of the effectiveness of NBW we needed to compare the retrieval performance of NBW with the unexpanded retrieval. To this end the popular baseline ranking model of BM25 (Robertson *et al.*, 1998; Carpineto *et al.*, 2001) was selected.

In the first-pass retrieval the baseline ranking model was used to retrieve documents for the query. Before the second-pass retrieval 10 of the retrieved documents with the highest rank were set as pseudo-relevant documents. Then NBW was used to weight the candidate expansion terms in the pseudo-relevant documents and was then combined with the Rocchio framework to re-weight the candidate and the original query terms. For the final weighting of the terms the 40 highest-ranked terms were selected to form a new query with which to run the second-pass retrieval. The baseline ranking model of BM25 was also used for the second-pass retrieval.

Table III presents a comparison of the results of the unexpanded query of BM25 as well as the expanded query with NBW applied. The measurements were primarily taken in terms of overall retrieval precision. For average precision NBW gained an increase of 0.05, a 12.2 percent increase rate. For R-precision NBW obtained an increase of 0.0468, an 11.79 percent increase rate. These two basic measurements show that NBW consistently achieved better retrieval effectiveness, showing increase rates of over 10 percent, compared to BM25 in terms of overall performance.

To give some insight into NBW's performance we present the measurements of average precision at various recall levels in Table IV. Notice that the values of the precision increase in the first three rows of the third column are especially low, below 0.0027 or even negative. The same situation occurs for the values of the increase rate in

Table III.
Retrieval effectiveness of AV-PREC and R-PREC for BM25and NBW

	BM25	NBW	Increase	Increase rate (%)
AV-PREC	0.4107	0.4608	0.0501	+12.20
R-PREC	0.3911	0.4379	0.0468	+11.79

Notes: $\alpha = 1$, $\beta = 1.5$

Table IV.
Retrieval effectiveness of AV-PREC for BM25 and NBW at various recall levels

	BM25	NBW	Increase	Increase rate (%)
Recall-0%	0.8242	0.8244	0.0002	+0.02
Recall-10%	0.7916	0.7900	−0.0016	−0.20
Recall-20%	0.6877	0.6904	0.0027	+0.39
Recall-30%	0.5719	0.6062	0.0343	+6.00
Recall-40%	0.4831	0.5391	0.056	+11.60
Recall-50%	0.4223	0.4833	0.061	+15.63
Recall-60%	0.3281	0.4097	0.0816	+24.87
Recall-70%	0.2314	0.3106	0.0792	+34.23
Recall-80%	0.1776	0.2582	0.0806	+45.38
Recall-90%	0.1263	0.1910	0.0647	+51.23
Recall-100%	0.1135	0.1674	0.0539	+47.49

the first three rows of the fourth column which are below 0.39 percent. However the values of the precision increase in the fifth through 11th rows of the third column are high and range from 0.0539 to 0.0816. Together the values of the increase rate in the fifth through 11th rows of the fourth column are also high and range from 11.6 percent to 51.23 percent.

Viewing these results together with the precision values of BM25 in the second column, the connection between the change of BM25's precision values and the change of the precision increase or the precision increase rate is noteworthy. As BM25's precision value approaches 0.7, the precision increases made by NBW are very limited. In contrast, as BM25's precision value is below 0.5, NBW can obtain a precision increase between 0.05 and 0.08 and a precision increase rate between 11 and 51 percent. These results reveal that NBW performed better when the performance of BM25 was lower or when the recall level was higher.

To gain a better perspective on NBW's performance we plotted the data into a two-dimensional space as shown in Figure 1. There are some performance characteristics worth noting. First as the recall levels are low and precision is high, NBW's performances are the same as BM25's. It seemed that NBW did not select new terms for query in these situations. Second NBW outperforms BM25 consistently when the latter's retrieval precision is lower. Third there is a clear tendency that NBW's performance follows BM25's. These characteristics indicate that NBW's performance was stable. In other words it indicates that NBW did not select expansion terms randomly. If NBW selected expansion terms randomly for query, the precisions derived for it would be randomly distributed. Therefore the results might not support NBW as being well developed, but could support the rejection of NBW as random.

Experiment 2

In this experiment we compared NBW with three alternate expansion term weighting functions compared by Carpineto *et al.* (2001). These included Doszkocs' function (CHI-1) of $weight_{CHI-1}(t) = [[P_R(t) - P_C(t)]/P_C(t)]$, the standard chi-square function (CHI-2) of $weight_{CHI-2}(t) = [[P_R(t) - P_C(t)]^2/P_C(t)]$, and the Carpineto *et al.* function (KLD) of equation 2. Carpineto *et al.* (2001) compared CHI-1, CHI-2, and KLD to show

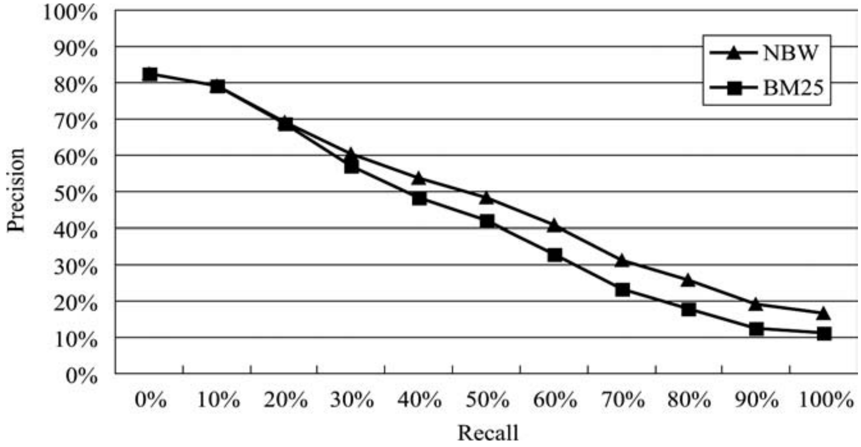

Figure 1.
Precision vs recall for BM25 and NBW

that the proposed KLD function could improve retrieval effectiveness. We applied Carpineto *et al.*'s (2001) method of comparison to evaluate how a different approach with the same probability base of NBW could perform in contrast to CHI-1, CHI-2, and KLD. The experimental procedures were the same as those of Experiment 1, but CHI-1, CHI-2, KLD, and NBW each served as the expansion term weighting function.

Table V presents the overall retrieval effectiveness measured in average precision and R-precision for the four expansion term weighting functions. It also compares each one to BM25 in terms of precision increase. Comparing each weighting function to BM25 fills two roles. The first is to measure the adequacy of the comparison between the four weighting functions. The results show that the four weighting functions all outperform the baseline ranking model (BM25) in both average precision and R-precision. These results ensure that the comparison between the expanding effectiveness for the four weighting functions is appropriate. The second is to compare the four weighting functions to the same base.

Comparing average precision the sequence order from low to high for the four weighting functions is: CHI-1 (an increase of 0.0057, a 1.39 percent increase rate); CHI-2 (an increase of 0.0274, a 6.67 percent increase rate); KLD (an increase of 0.0304, a 7.4 percent increase rate); and NBW (an increase of 0.0501, a 12.2 percent increase rate). Compared with R-precision the sequence order from low to high for the four weighting functions is the same: CHI-1 (an increase of 0.0057, a 1.46 percent increase rate); CHI-2 (an increase of 0.0225, a 5.57 percent increase rate); KLD (an increase of 0.0226, a 5.78 percent increase rate); and NBW (an increase of 0.0468, a 11.97 percent increase rate). For the three weighting functions of CHI-1, CHI-2, and KLD, the comparison results from this experiment were consistent with Carpineto *et al.*'s (2001) study in which CHI-1 was the worst while KLD was the best.

Since KLD is the best of the three, we thus make a direct comparison between NBW and KLD. For average precision NBW gains an increase of 0.0197, a 4.46 percent increase rate over KLD; for R-precision, NBW obtains an increase of 0.0242, a 5.85 percent increase rate over KLD. This precision increase can be contrasted with the gains of KLD over CHI-1 and CHI-2. For average precision KLD obtains an increase of 0.0247, a 5.93 percent increase rate over CHI-1 as well as an increase of 0.003, a 0.68

Table V.
Retrieval effectiveness of AV-PREC and R-PREC for each weighting function

	AV-PREC	R-PREC
BM25	0.4107	0.3911
CHI-1	0.4164	0.3968
Increase	0.0057	0.0057
Increase rate	+1.39	+1.46
CHI-2	0.4381	0.4136
Increase	0.0274	0.0225
Increase rate	+6.67	+5.75
KLD	0.4411	0.4137
Increase	0.0304	0.0226
Increase rate	+7.40	+5.78
NBW	0.4608	0.4379
Increase	0.0501	0.0468
Increase rate	+12.20	+11.97

percent increase rate over CHI-2; for R-precision, KLD obtains an increase of 0.0169, a 4.26 percent increase rate over CHI-1 as well as an increase of 0.0001, a 0.02 percent increase rate over CHI-2. These contrasts show that KLD's performance is about the same as CHI-2, while NBW's performance over KLD is similar to KLD's performance over CHI-1. Through this analysis we conclude that NBW has been developed enough to achieve acceptable performance.

Discussion

The test results above have presented some characteristics of NBW. First the parameter study on the value of β shows that the proposed weighting function could select usable expansion terms. Second the comparison of the retrieval effectiveness of NBW with BM25 has revealed performance tendencies indicating that NBW's performance is non-random and stable. Third, in terms of retrieval precision for achievement measurement, NBW's performance is acceptable in contrast to the other three weighting functions.

NBW was designed to be applied in any environment involving information retrieval, for example in searching webpages on a server, searching documents in a corporation's repository, or searching historical records in a hospital's database. Although the results seem to support the applicability of NBW there are limitations to be noted in the application of NBW or in a generalisation of this study's results. First NBW was implemented into Rocchio's (1971) framework and tested on query expansion. Its effectiveness in information retrieval can be claimed only within this framework and application. Second since the characteristics of the data set are not generalisable, more data sets and real life environments need to be tested before reaching generalisability. Third, since different baseline rankings can cause the weighting functions to perform differently, the test results produced on one baseline ranking model do not support claims that the same results would be reproduced on another. Therefore the other baseline ranking models that are applicable to NBW require validation. Fourth the values of the parameters selected for NBW apply to the experimental situation only. As the situation changes, the optimal values of the parameters will need to be reselected.

Although the descriptive statistics of the test results report that NBW is more effective than BM25 and the other weighting functions in terms of retrieval precision, its superiority cannot be ensured because those differences in effectiveness have not been tested for statistical significance. This study proposed a new approach for the development of the expansion term weighting function. With an aim to disclose the feasibility of the proposed approach, the tests focussed on the performance characteristics of the developed expansion term weighting function. To further verify the superiority of the retrieval effectiveness of NBW over BM25 and other weighting functions, tests of statistical significance are required with three considerations regarding the experimental design. First, since the characteristics of data sets are different, different data sets need to be employed for testing and comparison. Second, since different baseline ranking models can take different design approaches, each weighting function needs to select its own best fit baseline ranking model before the comparison of their retrieval effectiveness. Third, there are parameters which will affect the performance of the weighting function in its application to query expansion. The appropriate values of these parameters must be studied and selected for each weighting function.

Comparing the proposed approach in this study to the vector space model and the probabilistic model approaches, the proposed approach's advantages are clear. Since the vector space model approach (Rocchio, 1971; Alshaar, 2008) calculates the term's appearance difference between relevant and non-relevant documents as the basis for weighting candidate expansion terms, it does not reflect the term's real specialty in terms of its appearance in relevant documents. The probabilistic model (Harter, 1975a, b; Harman, 1992; Amati and van Rijsbergen, 2002; Alshaar, 2008) however measures and diversifies the term's appearance probability in relevant documents and the whole document set to reflect the term's real specialty in terms of its appearance in relevant documents. In spite of this advantage the probabilistic model approach has the disadvantage that the term's specialty revealed by different appearance situations such as relevance degree cannot be easily adjusted as in the vector space model approach. The development of NBW can take advantage of valuable aspects of both the vector space model and the probabilistic model approaches. To justify the feasibility of the proposed approach, NBW has been compared with the baseline ranking model to initially verify its stability. NBW has also been compared with Carpineto *et al.*'s (2001) probabilistic model – which had been shown to perform better than other probabilistic models – and Rocchio's (1971) vector space model. Despite the limitations of generalisation and deduction the study results have provided strong support for what this study has aimed to verify. Experimental results showing that the expansion term weighting function could perform stably and achieve acceptable effectiveness have highlighted the potential of this approach for further study and development.

Conclusion

Previous studies have used two important approaches toward the application of relevant/non-relevant information for query expansion. These two approaches have their own logic of information manipulation. This study proposed an interdisciplinary approach with valuable concepts taken from a combination of both approaches. The major contribution of this study is the exploitation of these concepts to produce an algorithm of an expansion term weighting function that performs effectively. This has created a new way of thinking about the development of the expansion term weighting function. In addition the proposed term weighting logic could be applied in an information retrieval study involving the weighting of terms in the vector space model for the creation of a user profile. Each of these could expand the field of culling information from documents.

The following are some advanced topics worthy of study:

- *An exploration of the limitations of NBW*. A data set used for an experiment can vary in subject, document number, query number, document length, query length, relevant document number, and so forth. How NBW will perform in a data set with different characteristics, such as TREC, requires examination. After testing alternate data sets, its performance in real working environments, such as on a Google search, can then be studied.

  Comparing NBW to other expansion term weighting functions requires a series of experimental designs where the characteristics of data and parameters, such as α and β, in the framework need to be controlled for study. Only with these series of comparisons will a researcher be able to determine the advantages and disadvantages of NBW against other weighting functions.

- *Further development of NBW.* In the NBW formula there are adjustments imposed on the term's probability value. We suggest that the design of these adjustments receive further study. One adjustment concerns the two axes' powers. We believe that the two axes' powers are not fixed. In a different situation, the emphasising values given to the two axes should be different. Advanced development of the formula could consider the setting of constants to adapt to different situations. Another adjustment that might lead to an improvement concerns document ranking. In NBW ranking is determined by a similarity comparison between the document and the query. One possible way to refine that adjustment is to cluster the pseudo-relevant documents by document similarity while working on the document ranking. This might enhance NBW's ability to distinguish relevant documents from non-relevant documents.
- *Application of the term weighting logic to the study of relevance feedback*. With relevance feedback the relevant and non-relevant document sets are provided. One application would create a user profile by selecting and weighting the terms from the document sets to assist further query. Since the fundamental process in the creation of the user profile is similar to the query expansion of our study, the term weighting logic of our design could undoubtedly be applied.

References

Alshaar, R. (2008), "Measuring the stability of query term collocations and using it in document ranking", Master's thesis, University of Waterloo, available at: http://hdl.handle.net/10012/4256 (accessed 31 July 2009).

Amati, G. and van Rijsbergen, C.J. (2002), "Probabilistic models of information retrieval based on measuring the divergence from randomness", *ACM Transactions on Information Systems*, Vol. 20 No. 4, pp. 357-89.

Attar, R. and Fraenkel, A.S. (1977), "Local feedback in full-text retrieval systems", *Journal of the ACM*, Vol. 24 No. 3, pp. 397-417.

Azimi-Sadjadi, M., Salazar, J., Srinivasan, S. and Sheedvash, S. (2004), "An adaptable connectionist text retrieval system with relevance feedback", *Proceedings of IEEE International Joint Conference on Neural Networks, Budapest*, IEEE, Washington, DC, pp. 309-14.

Balabanovic, M. (1997), "An adaptive webpage recommendation service", *Proceedings of the 1st International Conference on Autonomous Agents, Marina del Rey*, ACM Press, New York, NY, pp. 378-85.

Buckley, C., Salton, G., Allan, J. and Singhal, A. (1994), "Automatic query expansion using SMART", in Harman, D.K. (Ed.), *Proceedings of the 3rd Text Retrieval Conference (TREC-3)*, National Institute of Standards and Technology, Gaithersburg, MD, pp. 69-80.

Carpineto, C. and Romano, G. (2000), "Order-theoretical ranking", *Journal of the American Society for Information Science*, Vol. 51 No. 7, pp. 587-601.

Carpineto, C., Romano, G. and Giannini, V. (2002), "Improving retrieval feedback with multiple term-ranking function combination", *ACM Transactions on Information Systems*, Vol. 20 No. 3, pp. 259-90.

Carpineto, C., Mori, R.D., Romano, G. and Bigi, B. (2001), "An information-theoretic approach to automatic query expansion", *ACM Transactions on Information Systems*, Vol. 19 No. 1, pp. 1-27.

Chau, M., Fang, X. and Liu Sheng, R.O. (2005), "Analysis of the query logs of a website search engine", *Journal of the American Society for Information Science*, Vol. 56 No. 13, pp. 1363-7.

Chirita, P.A., Firan, C.S. and Nejdl, W. (2007), "Personalized query expansion for the web", *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam*, ACM Press, New York, NY, pp. 7-14.

Cover, T.M. and Thomas, J.A. (1991), *Elements of Information Theory*, Wiley-Interscience, New York, NY.

Croft, W.B. and Harper, D.J. (1979), "Using probabilistic models of document retrieval without relevance information", *Journal of Documentation*, Vol. 35 No. 4, pp. 285-95.

Cui, H., Wen, J.R., Nie, J.Y. and Ma, W.Y. (2003), "Query expansion by mining user logs", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15 No. 4, pp. 829-39.

Deerwester, S., Dumai, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990), "Indexing by latent semantic analysis", *Journal of the American Society for Information Science*, Vol. 41 No. 6, pp. 391-407.

Doszkocs, T.E. (1978), "AID: an associative interactive dictionary for online searching", *Online Information Review*, Vol. 2 No. 2, pp. 163-73.

Efthimiadis, E. (1993), "A user-centered evaluation of ranking algorithms for interactive query expansion", in Korfhage, R., Rasmussen, E. and Willett, P. (Eds), *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in Pittsburgh*, ACM Press, New York, NY, pp. 146-59.

Efthimiadis, E. (1995), "User choices: a new yardstick for the evaluation of ranking algorithms for interactive query expansion", *Information Processing and Management*, Vol. 32 No. 4, pp. 605-20.

Efthimiadis, E. and Biron, P. (1994), "UCLA-Okapi at TREC-2: query expansion experiments", in Harman, D.K. (Ed.), *Proceedings of the 2nd Text Retrieval Conference (TREC-2), National Institute of Standards and Technology, Gaithersburg, MD*, pp. 279-90.

Evans, D. and Lefferts, R. (1994), "Design and evaluation of the CLARITTREC-2 system", *Proceedings of the 2nd Text Retrieval Conference (TREC-2), National Institute of Standards and Technology, Gaithersburg, MD*, pp. 137-50.

Furnas, G.W., Deerwester, S., Dumais, S.T., Landauer, T.K., Harshman, R.A., Streeter, L.A. and Lochbaum, K.E. (1988), "Information retrieval using a singular value decomposition model of latent semantic structure", in Chiaramella, Y. (Ed.), *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Grenoble*, ACM Press, New York, NY, pp. 465-80.

Hariri, N. (2008), "An investigation of the effectiveness of the 'similar pages' feature of Google", *Online Information Review*, Vol. 32 No. 3, pp. 370-8.

Harman, D. (1992), "Relevance feedback revisited", in Belkin, N., Ingwersen, P. and Pejtersen, A.M. (Eds), *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen*, ACM Press, New York, NY, pp. 1-10.

Harter, S.P. (1975a), "A probabilistic approach to automatic keyword indexing. Part I: On the distribution of specialty words in a technical literature", *Journal of the American Society of Information Science*, Vol. 26 No. 4, pp. 197-206.

Harter, S.P. (1975b), "A probabilistic approach to automatic keyword indexing. Part II: an algorithm for probabilistic indexing", *Journal of the American Society of Information Science*, Vol. 26 No. 5, pp. 280-9.

Jing, Y. and Croft, W.B. (1994), "An association thesaurus for information retrieval", *Proceedings of RIAO'94: Intelligent Multimedia Information Retrieval Systems and Management, New York, 11-13 October*, CID, Paris, pp. 146-60.

Kim, B.M., Kim, J.Y. and Kim, J. (2001), "Query term expansion and re-weighting using term co-occurrence similarity and fuzzy inference", *Proceedings of IFSA World Congress and the 20th NAFIPS International Conference, Vancouver*, IEEE Standards Office, New York, NY, pp. 715-20.

Koster, C.H. and Beney, J.G. (2007), "On the importance of parameter tuning in text categorization", *Lecture Notes in Computer Science*, No. 4378, pp. 270-83.

Losee, R.M. (1990), *The Science of Information: Measurements and Application*, Academic Press, San Diego, CA.

Manning, C.D., Raghavan, P. and Schütze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press, New York, NY.

Mitra, M., Singhal, A. and Buckley, C. (1998), "Improving automatic query expansion", *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne*, ACM Press, New York, NY, pp. 206-14.

Moschitti, A. (2003), "A study on optimal parameter tuning for Rocchio text classifier", *Lecture Notes in Computer Science*, No. 5075, pp. 546-7.

Nick, Z.Z. and Themis, P. (2001), "Web search using a genetic algorithm", *IEEE Internet Computing*, Vol. 5 No. 2, pp. 18-26.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C. and Johnson, D. (2005), "Terrier information retrieval platform", *Lecture Notes in Computer Science*, No. 3405, pp. 517-19.

Porter, M. and Galpin, V. (1988), "Relevance feedback in a public access catalogue for a research library: Muscat at the Scott Polar Research Institute", *Program: Electronic Library and Information Systems*, Vol. 22 No. 1, pp. 1-20.

Qiu, Y. and Frei, H.P. (1993), "Concept-based query expansion", in Korfhage, R., Rasmussen, E. and Willett, P. (Eds), *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh*, ACM Press, New York, NY, pp. 160-9.

Robertson, S.E. and Spärck Jones, K. (1976), "Relevance weighting of search terms", *Journal of the American Society for Information Science*, Vol. 27 No. 3, pp. 129-46.

Robertson, S.E., Walker, S. and Beaulieu, M. (1998), "Okapi at TREC-7: automatic *ad hoc*, filtering, VLC, and interactive track", in Voorhees, E.M. and Harman, D.K. (Eds), *Proceedings of the 7th Text Retrieval Conference (TREC-7), National Institute of Standards and Technology, Gaithersburg, MD*, pp. 253-64.

Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M. and Gatford, M. (1995), "Okapi at TREC-3", in Harman, D.K. (Ed.), *Proceedings of the 3rd Text Retrieval Conference (TREC-3), National Institute of Standards and Technology, Gaithersburg, MD*, pp. 109-26.

Rocchio, J. (1971), "Relevance feedback in information retrieval", in Salton, G. (Ed.), *The SMART Retrieval System – Experiments in Automatic Document Processing*, Prentice-Hall, Englewood Cliffs, NJ, pp. 313-23.

Schütze, H. and Pedersen, J.O. (1994), "A co-occurrence-based thesaurus and two applications to information retrieval", *Information Processing and Management*, Vol. 33 No. 3, pp. 307-18.

Shanfeng, Z., Xiaotie, D., Kang, C. and Weimin, Z. (2001), "Using online relevance feedback to build effective personalized metasearch engine", in Özsu, M.T., Schek, H.T., Tanaka, H., Zhang, Y. and Kambayashi, Y. (Eds), *Proceedings of the 2nd Conference on Web Information Systems Engineering, Kyoto*, IEEE Computer Society, Washington, DC, pp. 262-8.

Singhal, A., Choi, J., Hindle, D., Lewis, D. and Pereira, F. (1999), "AT&T at TREC-7", in Voorhees, E.M. and Harman, D.K. (Eds), *Proceedings of the 7th Text Retrieval Conference (TREC-7), National Institute of Standards and Technology, Gaithersburg, MD*, pp. 239-52.

Spärck Jones, K. (1971), *Automatic Keyword Classification for Information Retrieval*, Butterworth, London.

Srinivasan, P. (1996), "Query expansion and MEDLINE", *Information Processing and Management*, Vol. 32 No. 4, pp. 431-43.

Xu, J. and Croft, W.B. (1996), "Query expansion using local and global document analysis", *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich*, ACM Press, New York, NY, pp. 4-11.

Xu, J. and Croft, W.B. (2000), "Improving the effectiveness of information retrieval with local context analysis", *ACM Transactions on Information Systems*, Vol. 18 No. 1, pp. 79-112.

Yu, C.T., Luk, W.S. and Cheung, T.Y. (1976), "A statistical model for relevance feedback in information retrieval", *Journal of the ACM*, Vol. 23 No. 2, pp. 273-86.

About the authors

Shihchieh Chou is an Associate Professor in the Department of Information Management at National Central University in Taiwan. He also serves as the Director of the Computer Centre at the Business School. He received his PhD degree from Texas A&M University in 1984. His research interests include information retrieval and knowledge management. He is the patent holder for two knowledge management inventions. Shihchieh Chou is the corresponding author and can be contacted at: scchou@mgt.ncu.edu.tw

Chinyi Cheng is a PhD Candidate in the Department of Information Management at National Central University. His research interests include information retrieval and knowledge management.

Szujui Huang is currently an Engineer at Inventec Corporation in Taiwan. He graduated from National Central University with a Master's degree. His interest in the study of information retrieval, knowledge management and ERP dates back to his student programme.