Online Information Review

Construction and application of specialty-term information for document re-ranking

Shihchieh Chou Zhangting Dai

Article information:

To cite this document:
Shihchieh Chou Zhangting Dai , (2016),"Construction and application of specialty-term information for document re-ranking", Online Information Review, Vol. 40 Iss 7 pp. 1054 - 1070
Permanent link to this document:
http://dx.doi.org/10.1108/OIR-04-2015-0128

Received 20 August 2015
Revised 21 March 2016
Accepted 12 May 2016

Construction and application of specialty-term information for document re-ranking

Shihchieh Chou and Zhangting Dai
Department of Information Management,
National Central University, Chung-Li, Taiwan

Abstract

Purpose – Conventional studies mainly classify a term's appearance in the retrieved documents as either relevant or irrelevant for application. The purpose of this paper is to differentiate the term's appearances in the retrieved documents in more detailed situations to generate relevance information and demonstrate the applicability of the derived information in combination with current methods of query expansion.
Design/methodology/approach – A method was designed first to utilize the derived information owing to term appearance differentiation within a conventional query expansion approach that has been proven as an effective technology in the enhancement of information retrieval. Then, an information retrieval system was developed to demonstrate the realization and sustain the study of the method. Formal tests were conducted to examine the distinguishing capability of the proposed information utilized in the method.
Findings – The experimental results show that substantial differences in performances can be achieved between the proposed method and the conventional query expansion method alone.
Practical implications – Since the proposed information resides at the bottom of the information hierarchy of relevance feedback, any technology regarding the application of relevance feedback information could consider the utilization of this piece of information.
Originality/value – The importance of the study is the disclosure of the applicability of the proposed information beyond current usage of term appearances in relevant/irrelevant documents and the initiation of a query expansion technology in the application of this information.

Keywords Query expansion, Information retrieval, Relevance feedback, Term appearance,
Term weight modification

Paper type Research paper

Introduction

Relevance feedback has been an important source for providing valuable information on the approach which aims to enhance information retrieval. In the past, the successful technology of query expansion and many studies on different methods have attested to its usability and wide applicability.

In the application of relevance feedback, the information on term appearances in relevant/irrelevant documents has been important and widely utilized. In the past, both models of query expansion study, i.e. the vector space model and the probabilistic model, have successfully applied this information for the enhancement of information retrieval. In terms of frequency or probability, the hypothesis is that the terms with query relevance differ from the terms without query relevance in their self-presentation in the relevant/irrelevant documents. The fundamental goal of the related studies is to

This work was supported in part by the National Science Council, Taiwan, under the Grant No. NSC 101-2410-H-008-001.

Online Information Review
Vol. 40 No. 7, 2016
pp. 1054-1070
© Emerald Group Publishing Limited
1468-4527
DOI 10.1108/OIR-04-2015-0128

measure the deviation of term appearance in relevant/irrelevant documents as the base for query term selection.

In the dealing with the feedback information as aforementioned, conventionally, the classification on documents is done first, and then the presence/absence of the terms in the classified documents is measured. In this paradigm, the term appearance situation is determined by the classification of documents as relevant or irrelevant. If we step further and consider a more detailed differentiation, a term could appear in one or more of the following document sets: first, relevant documents (abbreviated as an R term); second, irrelevant documents (abbreviated as an IR term); third, both relevant and irrelevant documents (abbreviated as an RIR term); fourth, relevant documents only (abbreviated as an RO term); and fifth, irrelevant documents only (abbreviated as an IRO term). For the five types of information, our interest is directed to the RO term in regard to the study of query expansion, owing to its special appearance characteristics, as with relevance and extreme deviation. Chou and Chang's (2009) study confirmed this specialty-term's distinguishing characteristics, and showed that it could be independently manipulated and applied in the enhancement of information retrieval.

The application of relevance feedback

Relevance feedback, which is one of the most popular ways for supporting relevance generation, refers to the judgment of a document's relevance after its initial retrieval. The two basic types of relevance feedback are interactive feedback and automatic feedback. In interactive feedback, the human process is needed in relevance identification for the retrieved documents. In automatic feedback, the top-ranked documents in the initial retrieval are viewed as relevant ones with pseudo relevance (Baeza-Yates and Ribeiro-Neto, 1999; Carpineto *et al.*, 2001; Manning *et al.*, 2008). In addition to these two types of relevance generation, there are also others, such as implicit relevance (Shen *et al.*, 2005), or clustering documents into different levels of relevance (Pu and He, 2009; Rooney *et al.*, 2006). Performing either one of the above methods after the initial retrieval, the relevance/irrelevance information could be set for extraction and application.

In the past, query expansion was one of the key approaches in the application of relevance/irrelevance information in the enhancement of information retrieval. The basic idea of query expansion is to select terms from the feedback documents to reformulate the query such that the new query will be moved toward the relevant documents and away from the irrelevant ones. Studies have shown that the technology of query expansion can significantly improve the effectiveness of information retrieval. There are two main models employed in the approach of query expansion: the vector space model and the probabilistic model (Alshaar, 2009; Baeza-Yates and Ribeiro-Neto, 1999; Carpineto *et al.*, 2001; Harman, 1992; Manning *et al.*, 2008).

In the vector space model, the main concern is the information of term appearance frequencies in the relevant/irrelevant documents (Harman, 1992; Alshaar, 2009). The kernel operation is to re-weight the terms of the query vector by adding the weights of the terms appearing in relevant documents and subtracting the weights of the terms appearing in irrelevant documents. With the incremental re-weighting of the relevant and irrelevant terms, the query vector will move toward the relevant documents and away from the irrelevant ones. Rocchio's (1971) method was one of the most famous and successful methods. Also, lots of variations with the similar idea had achieved good performances: Ide (1971) used the top irrelevant document only for feedback; Singhal *et al.* (1997) indicated that better results could be obtained by using documents close to the query of interest only, rather than all documents; Desjardins and Godin (2000) presented the development of a different weighting formula; Nick and Themis' (2001) approach was to rank the rating of the relevance of the retrieved documents; Kim *et al.*'s (2001) study calculated the relevance degree; Azimi-Sadjadi *et al.* (2007) study exploited relevance feedback from multiple expert users; Koster and Beney (2007) proposed the modification of parameters in Rocchio's original formula.

In the probabilistic model, the information of term appearance probabilities in the relevant/irrelevant documents is the main concern; numerous such studies were well performed, following Harter's (1975a, b) two-poisson model which was one of the very first to take this approach. Robertson and Sparck Jones (1976) proposed the classic probabilistic model known as binary independence retrieval. Harper and van Rijsbergen (1978) exploited the maximum spanning tree to form a term-term clustering technology for probabilistic query expansion. Wu and Salton (1981) utilized relevance feedback information to reweight terms with a probabilistic formula and deploy these terms for query expansion. Croft (1983) extended the weighting scheme by adapting the probabilistic formula to include "within document" frequency weights. Porter and Galpin's (1988) simple formula calculated the competing scores for the candidate expansion term. Harman (1992) revisited several important studies and reported over a decade's achievements of major works and study foci of the probabilistic approach.

In addition to the technology of query expansion, in recent years, there have been many studies on different techniques utilizing the aforementioned information of term relevance to enhance information retrieval. Takano *et al.*'s (2009) study applied the feedback information of classified documents to the creation of the vector space; to optimize the document vector space, the study developed a method to perform the operation of adding, subtracting and weighting of the relevant/irrelevant terms contained in the vector space. Chen and Lu's (2010) study converted the relevance ranking of the retrieved document into a document classification problem; they utilized the information of relevance derived from relevance feedback to train the SVM classifier in the categorization of relevant and irrelevant documents. Dang *et al.*'s (2014) study computed context-dependent term weights based on a mixture of unigrams and bigrams to improve the performance of relevance feedback retrieval. In the calculation of the new term weights, estimation of the local probability of relevance of each query term was required. Miao *et al.* (2012) utilized proximity-based term frequency in pseudo relevant documents to perform query expansion, having incorporated the proximity information into Rocchio's model. Haiduc *et al.* (2013) proposed exploiting the sample of queries and relevant results to train the classifier and use the classifier to recommend the best reformulation technique for the incoming queries. Vargas *et al.* (2013) proposed a method to select diverse expansion terms from a suitable partition of relevance feedback to tailor query expansion to the search goal. Rodriguez Perez *et al.* (2013) developed a re-ranking method by applying inter-document relations in relevance feedback. Dalton and Dietz (2013) proposed a neighborhood relevance model which utilized cross-document evidence in relevance feedback to identify the salient context.

Research question and purpose

The above studies present the applicability of relevance feedback, and that in the application of relevance feedback, the information of term appearance situations is important and worthy of study and examination in depth. In the dealing with the term appearance situation, conventional studies usually classify the terms into appearing in relevant or irrelevant documents as determined by relevance feedback. Our proposition is that the term appearance situation could be differentiated as R, IR, RIR, RO and IRO. Consider the appearance deviation of the RO term, its degree of relevance to the user's query interest could be higher than the terms of other appearance situations. An interesting question is: "Could the information of the RO term be applied beyond the current usage of the relevant/irrelevant information in query expansion for the enhancement of information retrieval?" To explore the application potential of the information of the RO term in query expansion, this study is concerned with the development and evaluation of a method that will make use of the information of the RO term in the enhancement of information retrieval, while retaining the current method of query expansion in its normal usage of the relevant/irrelevant information. In this way, we would be able to demonstrate that the information of the RO term is valuable in regard to the improvement of the current method of query expansion.

The method

Analysis

In the searching of relevant documents, a typical way is to compare the "similarity" between the query and the document. In this approach, both the query and the document are represented by a series of terms with term importance given to each term. When two sets of terms with term importance are identical, the query and the document represented by the two sets of terms are said to be totally similar, and the document is said to be completely relevant to the user's query interest. In the application of relevance feedback for query expansion in the vector space model, the main concern is the construction of a new query that will be more similar to the relevant documents. With this make up, the relevant document can be accessed just by comparing the similarity of the two term sets of the query and the document. In this approach, the terms' appearance frequencies in relevant/irrelevant documents together determine the term's importance (weight). The manipulation basically is to add the weight of the term as it appears in relevant documents and subtract the weight of the term as it appears in irrelevant documents. The selection of terms for query expansion is determined by the term's weight.

In this process, the two situations of "a term that appears only in relevant documents ten times" and "a term that appears in relevant documents 20 times and irrelevant documents ten times" have been equally treated to determine the term's weight. However, considering the appearance specialty, it seems reasonable to manipulate RO terms in a different way in the determination of the term's weight. Referring to the phenomenon of RO terms, there are four possibilities. First, the occurrence of RO is meaningful, and will continue to occur. In the further retrieved documents, the one comprising RO terms is certainly relevant. Second, the occurrence of RO is meaningful, but may stop occurring by chance. In the further retrieved documents, the ones comprising RO terms could be relevant or irrelevant. Third, the occurrence of RO is by chance, but will continue to occur by chance. In the further retrieved documents, the ones comprising RO terms will be relevant. Fourth, the occurrence of RO is by chance, and will stop occurring. In the further retrieved documents, the ones comprising RO terms could be relevant or irrelevant. For the four possibilities, will they happen evenly? Based on the phenomenon of appearance deviation, our deduction is that the occurrence ratio for the first and second possibilities should be higher than for the third and fourth; and in the second possibility, the RO terms to appear in relevant documents should be more numerous than in the irrelevant documents. If this deduction is correct, the appearances of RO terms in relevant documents should be more numerous than in irrelevant documents in the further retrieved documents. Thus, if we implement a strategy to increase the importance of RO terms in the construction of a new query, it would be beneficial to the retrieving of relevant documents. Since past query expansion studies did not take the applicability of RO terms into consideration, imposing this strategy on the conventional query expansion method, e.g. Rocchio's, would be able to verify whether the information of RO terms could be applied beyond the current usage of the relevant/irrelevant information in query expansion in the enhancement of information retrieval.

Embodiment

In this study, we propose a method called "Construction and application of specialty-term information" (CASTI) which will have the terms that appear in relevant documents only been identified and applied in information retrieval, while retaining the conventional classification of term appearance situations in its normal usage. It comprises the following four steps: term appearance identification, query expansion, expanded query modification and document re-ranking. Figure 1 shows the flow of using CASTI.

At the beginning, the user searches documents with the initial query. In relevance feedback, the retrieved documents are reviewed and judged by the user to form a sorted list on query relevance. After relevance feedback, CASTI starts. In the step of term appearance identification, three term sets will be constructed for application according to its appearance situation in the feedback documents, including: terms appearing in relevant documents (R), terms appearing in irrelevant document (IR) and terms appearing in relevant documents only (RO). Each of the three term sets has its own characteristics in regard to document relevance.

In the step of query expansion, the R and IR term sets are used as usual to form an expanded query to enable the vector space of the query to move toward the relevant document. All methods of this sort developed in the past can be merged into CASTI at this step to test the proposition of this study, that the newly identified term appearance situation of RO could be further applied beyond the conventional application of relevant/irrelevant information. Since Rocchio's method has been deemed as the most

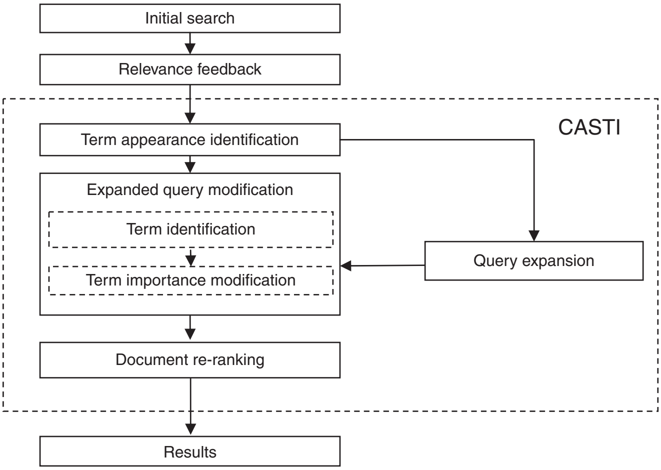


Figure 1.
The flow of using CASTI

popular and fundamental in the study of query expansion, this research will experiment with it to demonstrate the applicability of the proposed information.

In the conventional method of query expansion, a query term's importance is typically determined by adding the weight of the term as it appears in the relevant documents, and subtracting the weight of the term as it appears in the irrelevant documents. In these operations, the RO term's appearance specialty is not identified and treated. Therefore, the next step in CASTI, expanded query modification, is to have the information of RO terms applied in the modification of the expanded query constructed in the previous step to form a new query. The two sub-steps of "term identification" and "term importance modification" are performed sequentially to accomplish the job, as Figure 2 shows. In the sub-step of term identification, each term of the expanded query is checked with its existence or absence in the RO term set. Those terms in the expanded query which are identified as RO terms are marked for term importance modification in the next sub-step. In the sub-step of term importance modification, the weight of the marked term in the expanded query is modified by multiplying a value of *SI* as following equation shows, while keeping the original weight for the unmarked term. (See the appendix for the detailed computation of the two sub-steps.) Since the value of *SI* in following equation is utilized to increase the weight of the marked term or keep the original weight for the unmarked term, its value is set as greater than or equal to 1, and its best value for the marked term will be determined by the experiment of parameter study conducted later:

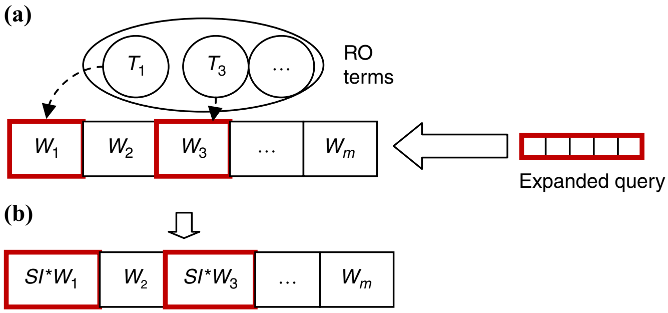
$$w_{q'j} = w_{qj} \times SI, \; SI \geqslant 1 \tag{1}$$

In the step of document re-ranking, the new query is utilized to re-rank the retrieved documents by computing similarities between the new query and each of the documents. In the process, term frequency and inverted document frequency are exploited to build the term weight of the document for similarity computation (Baeza-Yates and Ribeiro-Neto, 1999). All documents are re-ranked by similarity decrease for retrieval performance evaluation.

Experiments

Experimental parameters and settings

Two experiments were conducted in this study. The first experiment was performed to determine the best value of *SI* for CASTI. The second experiment compared the retrieval effectiveness of CASTI with Rocchio's method and the initial query.



Notes: (a) Term identification; (b) term importance modification

Figure 2. The step of expanded query modification in CASTI

Lemur had been adopted as the platform to implement the related system work for the experiments. We utilized the expanded query of Rocchio's method generated by Lemur's platform to develop CASTI and perform re-ranking. TREC6 was used to provide the required experimental data. In total, 35 topics of TREC6 were selected and utilized as the experimental data set based on the condition that the number of relevant documents in the initial retrieval should be greater than 20 to fulfill the demand of simulating relevance feedback.

In conducting the two experiments, we first used the title query in the selected topic as the initial query to simulate the user's initial retrieval; the initially retrieved documents in each topic were then judged as relevant or irrelevant by comparing them with the judgment file of each topic. The top 20 ranked relevant documents and the top 20 ranked irrelevant documents were used to simulate the user's relevance feedback. The rest of the documents in each topic were utilized in the experiments as the retrieved document set, with which document re-ranking was later performed by the three retrieval methods. All global parameters are listed in Table I.

The measurements on the experimental results included precision, mean average precision (MAP), normalized discounted cumulative gain (NDCG), NDCG15, Precision-recall and precision at N (P@N). Precision and MAP are measurements to assess the relevance of documents answered by a retrieval system. Precision measures the proportion of correct answers among the answers found by a retrieval system. MAP for a set of queries is the mean of the average precision scores for queries. These two measurements together are employed in this study to evaluate the overall performance on relevance retrieval. In addition to relevance, the other important retrieval performance of our concern is the ranking order for relevant documents. Poor ranking order for the list of retrieved documents with high precision of relevance still does not fit the requirements of an efficient search. This study utilizes the two measurements of

Parameter	Setting
Lemur version	4.12
Retrieval model	Vector space model
Similarity function	Cosine
Document collection	TREC6 (CD4 & 5)
Retrieved document range	The Documents listed in judgment file of each topic
Topics	301, 302, 304, 305, 306, 307, 311, 313, 314, 315, 316, 318, 319, 321, 322, 323, 324, 325, 326, 329, 330, 331, 332, 333, 335, 337, 340, 341, 342, 343, 345, 346, 347, 349, 350
The number of retrieved documents	1,000
The number of feedback relevant documents	20
The number of feedback irrelevant documents	20
The number of the documents eliminated feedback	960
Extended terms of Rocchio	All terms
α of Rocchio	1
β of Rocchio	1
γ of Rocchio	0
SI	1, 5, 10, 50, 100, 500, 1,000

Table I.
Parameter settings for the development CASTI

NDCG and NDCG15 to evaluate the performances of the system on relevance while considering ranking order. NDCG is a normalization of discounted cumulative gain (DCG), and DCG is a measure of ranking quality by measuring the usefulness (gain) of the document based on its relevance and position in the ranked list of the retrieved documents. NDCG15 is a version of NDCG for top-15 documents in the ranked list of the retrieved documents. Another important measurement to be considered in the evaluation of relevance is recall. While precision is the fraction of retrieved documents that are relevant, recall measures the proportion of relevant documents over the total relevant documents. The measure of recall is often used in combination with the measure of precision to form the analysis of precision-recall, in which either value for one measure is compared for a fixed level at the other measure. Often, there is an inverse relationship between precision and recall. With the analysis of precision-recall, we have aimed to present the performance tendencies of CASTI and the other two query methods. One more measurement of interest about relevance evaluation is on top N documents in the ranked list. Since the user's searching interest usually falls within the top N documents of the ranked list, the precision of relevance for the top *N* documents need to be concerned. Thus, P@N is employed in this study to measure the proportion of the relevant documents over the retrieved top N documents in the ranked list (Baeza-Yates and Ribeiro-Neto, 1999; Manning *et al.*, 2008).

Data presentation and analysis

The selection of the SI value. To detect the effect of the *SI* value, CASTI performed document re-ranking for each topic with different settings of *SI* values. Table II presents its average retrieval performances. The number of document sets on average is 956.6, the number of relevant documents on average is 127.6571 and the number of retrieved relevant documents on average is 88.8852.

In Table II, the situation worth noting is the changing of the *SI* value from 1 to 5, where all measurements except P@960 have gained great increases, ranging from 11 percent to 107 percent. As for the changing of *SI* values from five to ten and greater, the increase rates are small and not clearly evident. Figure 3(a) presents the performances for different P@Ns. As the figure shows, except for P@960 where no

	1	5	10	50	100	500	1,000
num_ret	956.6						
num_rel	127.6571						
num_rel_ret	88.8857						
Map	0.1380	0.1964 (+42%)	0.2034	0.2066	0.2071	0.2074	0.2074
NDCG	0.4388	0.4874 (+11%)	0.4923	0.4931	0.4934	0.4935	0.4934
NDCG15	0.2481	0.4310 (+74%)	0.4450	0.4524	0.4529	0.4517	0.4527
P5	0.2400	0.4971 (+107%)	0.5086	0.5086	0.5143	0.5143	0.5143
P10	0.2257	0.4571 (+103%)	0.4629	0.4686	0.4743	0.4686	0.4686
P15	0.2381	0.4038 (+70%)	0.4133	0.4248	0.4248	0.4229	0.4248
P20	0.2286	0.3757 (+64%)	0.3871	0.3986	0.4014	0.4043	0.4043
P30	0.2143	0.3505 (+64%)	0.3667	0.3724	0.3724	0.3714	0.3714
P100	0.1780	0.2491 (+40%)	0.2606	0.2669	0.2677	0.2674	0.2674
P200	0.1597	0.1984 (+24%)	0.2039	0.2077	0.2087	0.2094	0.2097
P500	0.1230	0.1375 (+12%)	0.1394	0.1410	0.1409	0.1410	0.1410
P960	0.0889	0.0889 (+0%)	0.0889	0.0889	0.0889	0.0889	0.0889

Table II. Performances of CASTI in different settings of *SI* values

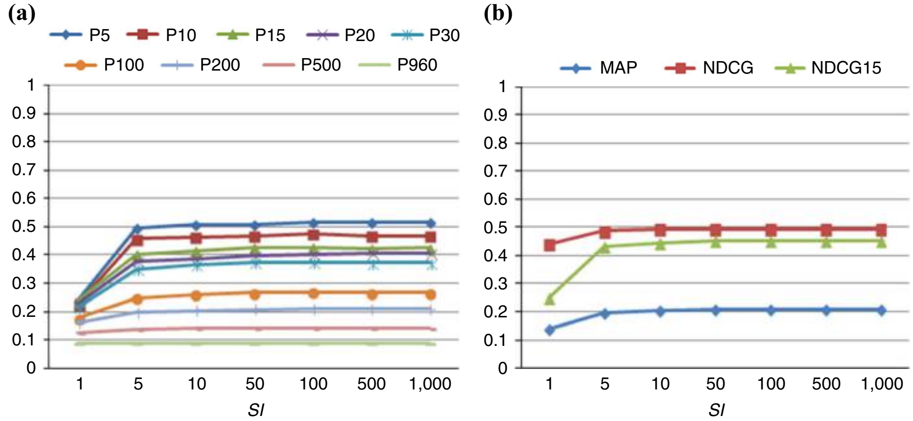


Figure 3. Performances of CASTI in different settings of *SI* values

difference in performance is presented, a sharp rise in performance can be seen for the other P@Ns when the *SI* value is changed from one to five. However, in the changing of *SI* values from 5 to 10 and greater, the increases in performances are not clear. Figure 3(b) presents the performances under the measurements of MAP, NDCG and NDCG15; the tendencies are consistent with Figure 3(a). Based on the results and analyses, this study thus adopted the value of 5 for *SI* in the implementation of CASTI.

Comparisons of CASTI, Rocchio's method and the initial query. In order to produce an overall view of the retrieval effectiveness of CASTI ($SI=5$), the retrieval performances of CASTI, Rocchio's method and the unexpanded initial query were measured and compared. Table III presents the retrieval performances measured in average precision, NDCG and NDCG15 for the three retrieval methods. We first compare Rocchio's method with the initial query which is set as the baseline. For average precision, Rocchio's method gained an increase of 0.0372, a 37 percent increase rate, over the initial query. For NDCG, which considers not only precision, but also the ranking order, Rocchio's method gained an increase of 0.0552, a 14 percent increase rate, over the initial query. For NDCG15, which is similar to the measurement of NDCG but considers the ranking order of the sorted top 15, Rocchio's method gained an increase of 0.1905, a 331 percent increase rate, over the initial query. The results show that, as expected, Rocchio's method can achieve better performances than the initial query. It attests that the adoption of Rocchio's method as the base in the development of CASTI is proper, and that it is adequate to compare the performances of CASTI with Rocchio's method in the evaluation of CASTI's effectiveness. As Table III shows, for average precision, CASTI gained an increase of 0.0584, a 42 percent increase rate, over Rocchio's method. For NDCG, CASTI gained an increase of 0.0486, an 11 percent increase rate, over Rocchio's method. For NDCG15, CASTI gained an increase of 0.1829,

Table III. Measurements of MAP, NDCG and NDCG15 for the three methods

	The initial query	Rocchio	CASTI
MAP	0.1008	0.1380	0.1964
NDCG	0.3835	0.4388	0.4874
NDCG15	0.0576	0.2481	0.4310

a 74 percent increase rate, over Rocchio's method. The results of these three basic measurements show that CASTI consistently achieves better retrieval performances compared with Rocchio's method in terms of overall performance.

To contrast the overall performing tendencies of the three methods, we present the measurements of precision-recalls in Figure 4. As Figure 4 shows, there is a clear tendency such that the three methods' performances follow each other, while both CASTI and Rocchio's method outperform the initial query, and CASTI outperforms Rocchio's method. These characteristics could support the rejection of CASTI's performance as merely random. It also indicates that CASTI is stable and well developed.

Since the top five or ten documents in the ranked list usually are the user's searching focus, it is of interest to compare the performances in terms of P@5 and P@10 for the three methods. As Table IV presents, for P@5, the precision order from low to high for the three methods is the initial query, Rocchio's method and CASTI, sequentially. Rocchio's method gained an increase of 0.1771 (a 282 percent increase rate) over the initial query, and CASTI obtained an increase of 0.2571 (a 107 percent increase rate) over Rocchio's method. For P@10, the results are consistent with P@5. The precision order from low to high for the three methods, again, is the initial query, Rocchio's method and CASTI, sequentially. Rocchio's method gained an increase of 0.1714 (a 316 percent increase rate) over the initial query, and CASTI obtained an increase of 0.2314 (a 103 percent increase rate) over Rocchio's method. The above results show that CASTI's performances in terms of P@5 and P@10 are consistent with the overall performances in outperforming Rocchio's method as aforementioned.

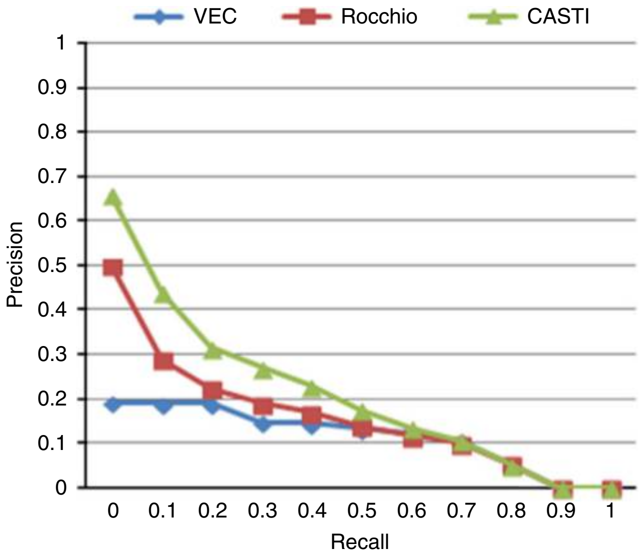


Figure 4.
Measurements of precision-recall for the three methods

Table IV.
Measurements of P@5 and P@10 for the three methods

	The initial query	Rocchio	CASTI
P@5	0.0629	0.2400	0.4971
P@10	0.0543	0.2257	0.4571

To have an inside perspective of the three methods' performances on every experimental topic under P@5 and P@10, we draw the two bar charts of Figures 5 and 6 to present the results. To compare the stability of the three methods under P@5, for each method, we counted the number of the topics that a method could rank documents in the top 5. The sequence order of the counted numbers for the three methods from low to high is: three for the initial query, 18 for Rocchio's method and 27 for CASTI. To compare the efficiency of the three methods under P@5, for each method, we counted the number of the topics that a method had achieved best performances. The sequence order of the counted numbers for the three methods from low to high is: two for the initial query, three for Rocchio's method and 26 for CASTI. Regarding P@10 with the abovementioned comparisons, the tendencies of the performances for the three methods are similar to P@5. In terms of the stability of the three methods, the sequence order of the counted numbers for the three methods from low to high is: four for the initial query, 23 for Rocchio's method and 28 for CASTI. In terms of the efficiency of the three methods, the sequence order of the counted numbers for the three methods from low to high is: two for the initial query, four for Rocchio's method and 26 for CASTI.

Comparisons of the above have presented three characteristics worth noting. First, the results are consistent with the average precision measurements. Second, CASTI's performance is evenly distributed. Third, it reveals that the measurement of high average precision for CASTI's performance is not caused by some extreme effects. To conclude, comparisons and analyses of the above support that CASTI is superior to Rocchio's method in stability and efficiency in terms of P@5 and P@10.

Discussions

The main purpose of the experiments was to examine the effect of applying the information of RO terms in combination with the current query expansion method. We compared the retrieval performances of CASTI with the query expansion method of Rocchio, and the results showed substantial differences in performances. An adequate interpretation of this result is that the information of RO terms has the potential to be utilized beyond the current application of relevant/irrelevant information by the query expansion method.

In regard to the possible generalization of the study's results, further studies of at least the following are needed. First, according to the experiments, the setting of the *SI* value affects the performance of CASTI; therefore, finding a way to determine the appropriate value of *SI* for CASTI in different application environments requires further study. Second, to further verify the superiority of the retrieval effectiveness of CASTI over the conventional query expansion method, tests of statistical significance are needed. Third, the study's results have been based on the data set of TREC; as the data set will be changed to yield different data characteristics or in the real world environment, the results could be different. Therefore, tests on verified data environments are required. Fourth, as there are query expansion methods with different designs of utilizing relevance feedback, will the information of RO terms work in those situations? This needs to be clarified. In concluding, before future studies would clarify the aforementioned matters, the claims of the proposed effect of the information of RO terms should be limited to this study only.

Implications

In the application of relevance feedback, the information like frequency, co-occurrence, or relevant/irrelevant classification of document terms are the basic information which

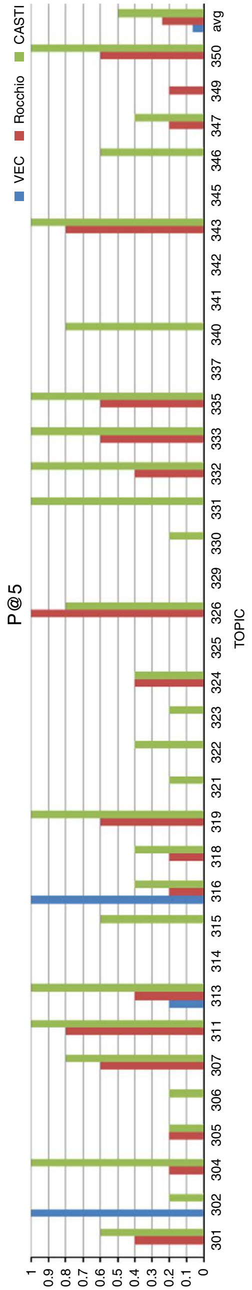


Figure 5.
Performances of the three methods for all topics under P@5

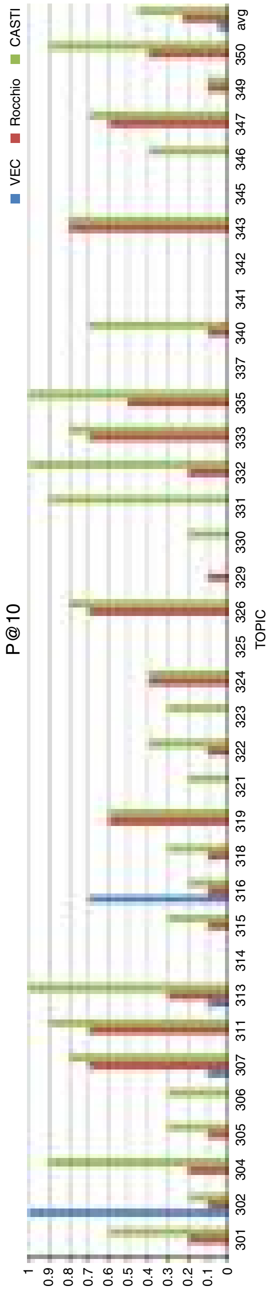


Figure 6.
Performances of the three methods for all topics under P@10

can be utilized to derive advanced information to support different kinds of information retrieval studies, e.g. building document vectors, classifying documents or constructing document concepts. The proposed information of RO terms could be classified as the mentioned type of basic information. Since the applicability of this basic information of RO terms has been shown in this research, the theoretical implication is that various kinds of studies having to do with the application of relevance feedback could extend their theoretical considerations to the proposed piece of information.

This study has demonstrated that the information of RO terms could be utilized in combination with Rocchio's method of query expansion to further enhance retrieval effectiveness. The practical implication lies in that Rocchio's query expansion method has been successfully tested over a long period of time, and it has been the most often adopted technique in the use of relevance feedback in achieving efficient search for large data sets. Since it is easy for the operational system to have the information of RO terms used together with Rocchio's query expansion technique, the information of RO terms could have a better chance to be tested in the operational environment through its residing in Rocchio's technique. If the retrieval effectiveness as mentioned could be further confirmed, the information of RO terms could obtain a wide usage in and present its contribution to the operational environment of query expansion.

Conclusion

Previous studies mainly classified the feedback documents into relevant and irrelevant for query expansion application. Whether in the vector space model or the probabilistic model, the manipulation of the relevant/irrelevant information was based on the relevance of the term to the query, as determined by checking the term's presence/absence in the relevant/irrelevant documents. Since the term's presence/absence in the relevant/irrelevant documents is of the most concern, this study proposes a more detailed differentiation of the term's appearance in documents. With this differentiation, the information related to "terms appearing in relevant documents only" emerges to reveal its application potential. The experiments' results show that the information of RO terms could be applied in combination with the current method of query expansion to further enhance retrieval effectiveness. The major contribution of this study is the disclosure of the applicability of the abovementioned piece of information, and the initiation of a query expansion technology in the application of this piece of information.

References

Alshaar, R. (2009), *Measuring the Stability of Query Term Collocations and Using it in Document Ranking*, University of Waterloo Library, Waterloo, available at: http://hdl.handle.net/10012/4256 (accessed July 31, 2009).

Azimi-Sadjadi, M.R., Salazar, J., Srinivasan, S. and Sheedvash, S. (2007), "An adaptable connectionist text-retrieval system with relevance feedback", *IEEE Transactions on Neural Networks*, Vol. 18 No. 6, pp. 1597-1613.

Baeza-Yates, R.A. and Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, Addison-Wesley Longman Publishing Co., Inc, Boston, MA.

Carpineto, C., de Mori, R., 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.

Chen, Z. and Lu, Y. (2010), "Using text classification method in relevance feedback", in Nguyen, N., Le, M. and Świątek, J. (Eds), *Intelligent Information and Database Systems*, Springer, Berlin and Heidelberg, pp. 441-449.

Chou, S. and Chang, W. (2009), "The identification of distinguishing term characteristics from relevance feedback", *Online Information Review*, Vol. 33 No. 4, pp. 745-760.

Croft, W.B. (1983), "Experiments with representation in a document-retrieval system", *Information Technology-Research Development Applications*, Vol. 2 No. 1, pp. 1-21.

Dalton, J. and Dietz, L. (2013), "A neighborhood relevance model for entity linking", *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, Le Centre de Hautes Etudes Internationales d" Informatique Documentaire*, *Lisbon and Paris*, pp. 149-156.

Dang, E.K.F., Luk, R.W.P. and Allan, J. (2014), "Beyond bag-of-words: bigram-enhanced context-dependent term weights", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 6, pp. 1134-1148.

Desjardins, G. and Godin, R. (2000), "Combining relevance feedback and genetic algorithm in an internet information filtering engine", *RIAO*, *College de France*, *Paris*, *April 12-14*, pp. 1676-1685.

Haiduc, S., Bavota, G., Marcus, A., Oliveto, R., Lucia, A.D. and Menzies, T. (2013), "Automatic query reformulations for text retrieval in software engineering", *Proceedings of the 2013 International Conference on Software Engineering, IEEE Press*, *San Francisco, CA, and Piscataway, NJ*, pp. 842-851.

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

Harper, D.J. and van Rijsbergen, C.J. (1978), "An evaluation of feedback in document retrieval using co-occurrence data", *Journal of Documentation*, Vol. 34 No. 3, pp. 189-216.

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 for 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 for Information Science*, Vol. 26 No. 5, pp. 280-289.

Ide, E. (1971), "New experiments in relevance feedback", in Salton, G. (Ed.), *The SMART Retrieval System*, Prentice Hall, Englewood Cliffs, NJ, pp. 337-354.

Kim, B.M., Kim, J.Y. and Kim, J. (2001), "Query term expansion and reweighting using term co-occurrence similarity and fuzzy inference", *IFSA World Congress and 20th NAFIPS International Conference, Joint 9th, IEEE*, *Vancouver*, *July 25-28*, pp. 715-720.

Koster, C.H.A. and Beney, J.G. (2007), "On the importance of parameter tuning in text categorization", in Virbitskaite, I. and Voronkov, A. (Eds), *Perspectives of Systems Informatics*, Springer, Berlin and Heidelberg, pp. 270-283.

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

Miao, J., Huang, J.X. and Ye, Z. (2012), "Proximity-based Rocchio's model for pseudo relevance", *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM*, *Portland, OR, and New York, NY*, pp. 535-544.

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

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*, Vol. 22 No. 1, pp. 1-20.

Pu, Q. and He, D. (2009), "Pseudo relevance feedback using semantic clustering in relevance language model", *Proceedings of the 18th ACM Conference on Information and Knowledge Management, ACM, Hong Kong, and New York, NY*, pp. 1931-1934.

Robertson, S.E. and Sparck Jones, K. (1976), "Relevance weighting of search terms", *Journal of the American Society for Information Science*, Vol. 27 No. 3, pp. 129-146.

Rocchio, J.J. (1971), "Relevance feedback in information retrieval", in Salton, G. (Ed.), *The SMART Retrieval System*, Prentice Hall, Englewood Cliffs, NJ, pp. 313-323.

Rodriguez Perez, J.A., Moshfeghi, Y. and Jose, J.M. (2013), "On using inter-document relations in microblog retrieval", *Proceedings of the 22nd International Conference on World Wide Web companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva*, *Rio de Janeiro*, pp. 75-76.

Rooney, N., Patterson, D., Galushka, M. and Dobrynin, V. (2006), "A relevance feedback mechanism for cluster-based retrieval", *Information Processing & Management*, Vol. 42 No. 5, pp. 1176-1184.

Shen, X., Tan, B. and Zhai, C. (2005), "Context-sensitive information retrieval using implicit feedback", *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM*, *Salvador, and New York, NY*, pp. 43-50.

Singhal, A., Mitra, M. and Buckley, C. (1997), "Learning routing queries in a query zone", *SIGIR Forum*, Vol. 31 No. SI, pp. 25-32.

Takano, K., Chen, X. and Masuda, K. (2009), "A framework for a feedback process to analyze and personalize a document vector space in a feature extraction model", *Information Technology and Management*, Vol. 10 Nos 2-3, pp. 151-176.

Vargas, S., Santos, R.L.T., Macdonald, C. and Ounis, I. (2013), "Selecting effective expansion terms for diversity", *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, Le Centre De Hautes Etudes Internationales d" Informatique Documentaire*, *Lisbon, and Paris*, pp. 69-76.

Wu, H. and Salton, G. (1981), "The estimation of term relevance weights using relevance feedback", *Journal of Documentation*, Vol. 37 No. 4, pp. 194-214.

Further reading

Salton, G. (1971), *The Smart Retrieval System: Experiments In Automatic Document Processing*, Prentice-Hall, Englewood Cliffs, NJ.

The Lemur Project. The Lemur Toolkit, University of Massachusetts and Carnegie Mellon University, available at: www.lemurproject.org/ (accessed October 25, 2012).

Appendix. Expanded query modification in CASTI

In CASTI, the step of expanded query modification is to have the information of RO terms applied in the modification of the expanded query constructed in the previous step to form a new query. The two sub-steps of "term identification" and "term importance modification" are performed sequentially to accomplish the job.

In the sub-step of term identification, each term (t_j) of the expanded query is checked with its existence or absence in the RO term set and marked as following Equation shows for term importance modification in the next sub-step:

$$F_t(t_j) = \begin{cases} \text{true}, & \text{if } t_j \text{ exists in RO} \\ \text{false}, & \text{if } t_j \text{ doesnot exist} \quad \text{in} \quad \text{RO} \end{cases} \tag{A1}$$

In the sub-step of term importance modification, the weights (w_{tj}) of all terms of the expanded query are used first to construct a unit matrix I_m:

$$I_m = \begin{bmatrix} w_{t1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{tm} \end{bmatrix}, w_{t1} = \cdots = w_{tj} = \cdots = w_{tm} = 1$$

then the weights of those terms (t_j) marked as true before are adjusted from 1 to *SI* as following equation presents, to form a term information matrix $M_{m \times m}$:

$$M_{m \times m} = F_m(I_m) = \begin{cases} w_{tj} = SI, \text{ if } F_t(t_j) = \text{true} \\ w_{tj} = 1, \text{ else} \end{cases} \tag{A2}$$

The term information matrix ($M_{m \times m}$) will be utilized to multiply the original query vector ($\vec{q} = \{w_{q1}, w_{q2}, \ldots, w_{qj}, \ldots, w_{qm}\}$) to form a new query vector ($\vec{q'} = \{w_{q'1}, w_{q'2}, \ldots, w_{q'j}, \ldots, w_{q'm}\}$), as Equation A3 shows. Both the original and the new query vectors are of dimension *m*. The content of each dimension of the original query vector contains weight (w_{qj}) created by the method developed in the past, such as Rocchio's. Those terms that have been marked as only appearing in relevant documents will have their weights in the dimension modified by multiplying the value of *SI,* as Equation A4 shows:

$$\vec{q'} = \vec{q} \times M \tag{A3}$$

$$w_{q'j} = w_{qj} \times SI, SI \geqslant 1 \tag{A4}$$

About the authors

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

Zhangting Dai is a PhD Candidate in the Department of Information Management at the National Central University in Taiwan. His research interests are in the fields of information retrieval, data mining and soft computing.