Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2014 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2014

THE APPLICATION OF SEMANTIC INFORMATION CONTAINED IN RELEVANCE FEEDBACK IN THE ENHANCEMENT OF DOCUMENT RE-RANKING

Shihchieh Chou
National Central University, scchou@mgt.ncu.edu.tw

Jiaxiong Zeng
National Central University, 94203009@cc.ncu.edu.tw

Zhangting Dai
National Central University, jkksdlinux@gmail.com

THE APPLICATION OF SEMANTIC INFORMATION CONTAINED IN RELEVANCE FEEDBACK IN THE ENHANCEMENT OF DOCUMENT RE-RANKING

Shihchieh Chou, Department of Information Management, National Central University, Taoyuan County, Taiwan, R.O.C., scchou@mgt.ncu.edu.tw

Jiaxiong Zeng, Department of Information Management, National Central University, Taoyuan County, Taiwan, R.O.C., 994203009@cc.ncu.edu.tw

Zhangting Dai, Department of Information Management, National Central University, Taoyuan County, Taiwan, R.O.C., jkksdlinux@gmail.com

Abstract

Easily accessed publishing channels have resulted in the problem of information overload. Conventional information retrieval models, such as the vector model or the probability model, apply the lexical information contained in relevance feedback in the enhancement of document re-ranking. Improvement is possible considering the application of semantic information. Studies have been taking the approach of concept extraction and application in the dealing with this semantic matter. So far, a perfect solution remains elusive and research still has new ground to cover. As such, we have proposed and tested a strategic method to form a more understanding of this field of study. The results of formal tests show that the proposed method is more effective than the baseline ranking model.

Keywords: Concept extraction, Concept fusion, Document re-ranking, Information retrieval, Semantic analysis, User's profile

1 INTRODUCTION

In conventional information retrieval, the lexical information contained in relevance feedback, such as term frequency or document frequency, is broadly utilized in the representation of the query, the document, and the user's profile. The information retrieval models taking this approach, such as the vector model or the probability model, are based on term matching for similarity measurement. Although this approach has made some success in the enhancement of information retrieval, improvements are still required.

Consider that a document selected in relevance feedback may consist of multiple topics, and not all topics are related to the query and suitable for application. The method of term matching could confront with this problem. The solution generally accepted is to extract the semantic information from terms for application. This work is not to solve the problems of polysemy and synonymy only. The main challenge is to differentiate between the lexical level ("what has been said or written") and the semantic level ("what was intended or what was referred to") in the text collection.

In the past, studies have been taking the approach of concept extraction and application in the dealing with this semantic matter. Although a lot of successful efforts have been made, a perfect solution remains elusive and research still has new ground to cover. In this study, our interest in the dealing with the semantic information is to extract the concept information from relevance feedback and apply this information in the enhancement of document re-ranking.

The rest of the paper is organized as follows. Section 2 contains related literatures for the work that we present in this paper. Section 3 details the method that we have developed for the extraction and application of concept information. Section 4 presents the evaluation of the method. Section 5 makes some conclusions and suggests some further works.

2 RELATED WORKS

This section presents some studies related to this research.

2.1 Term information

In modern information retrieval, query expansion is one of the most important approaches in the utilization of term information contained in relevance feedback in the enhancement of information retrieval. The most important two models are the Vector Space Model and the Probabilistic Model.

In the Vector Space Model, lots of studies can be exemplified: (1) Ranking of the relevance rating to the indexed terms. Each indexed term is attached a term weight formed by the term information, such as the term frequency (TF), or document frequency (DF). (Harman 1992; Alshaar 2008). Another main operation is to re-weight the terms of query vector according to the rule of adding the weights of terms in the relevant documents and subtracting the weights of terms in the irrelevant documents. Rocchio's method (Eq. 1) is one of the most famous, and lots of variations inspired with the idea also do well performance (Balabanovic et al. 1997; Nick & Themis 2001), (2) Refining a document ranking by learning filtering rule sets (Okabe et al. 2005), (3) Parameter modification in Rocchio's original formula (Yu et al. 1976; Koster et al. 2007; Moschitti 2003), (4) Combining the relevance feedback with the user preference and the ranking order (Shanfeng et al. 2001)

$$\overrightarrow{Q_m} = \alpha \overrightarrow{Q_0} + \beta \frac{1}{|D_r|} \sum_{\vec{d_j} \in D_r} \vec{d_j} - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d_j} \in D_{nr}} \vec{d_j} \quad (1)$$

In the Probabilistic Model, the term appearance probability is the main operation. Harter (1975a, b) proposed 2-poisson model (Amati & van Rijsbergen, 2002). Robertson and Sparck Jones (1976) explored the potential effectiveness of Harter's model for direct retrieval exploitation. Carpineto et al.(2001) gave an example of those studies in his formula of term weighting.

2.2 Semantic information

Although two sets of terms are semantically similar, the retrieval of it could fail because the two sets of terms are lexicographically different (Varelas et al. 2005). Undoubtedly, the application of semantic information is important and the technology of it has been widely developing. One important semantic technology is the Semantic Web, defined as: "an extension of the current Web in which information is given well-defined meaning" (Berners-Lee et al. 2001). Halliday and Hasan's study (1976) indicated that the semantic relatedness was calculable by utilizing the linguistic structure, lexical chains. In the SAFARI project, Shek et al. (1998) developed the method for content-based mining of image archives and integrated the semantic information to source matching. Moldovan and Mihalcea (2000) extracted the important information based on semantic meanings in query results to expand the proposed system and to improve the searches on the Web. Baziz et al. (2005) demonstrated the usage of ontology for information retrieval. The IntelliZap system (Finkelstein et al. 2002) built a client-server paradigm based on the information in user's marked terms for context search. Zhang et al. (2008) proposed a mechanism using the user's interested topic as query expansion. It utilizes the concept lattices (Wille 1992) to go through the domain terms needed by query expansion.

2.3 Semantic information models

The semantic information depicts the term relationships formed by the linguistic meanings. The term relationship is utilized as a latent variable in the important models of LSI (Deerwester et al. 1990), PLSA (Hofmann 1999), and LDA (Blei et al. 2003).

The latent semantic indexing (LSI) (Deerwester et al. 1990) identifies a linear subspace of TF-IDF matrix by singular value decomposition (Eq. 2) which represents the information compression in large collections. The main idea of LSI is to map high-dimensional vectors, such as the document vector in the vector space model, to a lower dimensional representation in a latent semantic space.

$$X = T_0 S_0 D_0 \tag{2}$$

PLSA is a statistical model based on the aspect model (Eq. 3) to automate document indexing utilizing the latent variable by the factor analysis of count term data. In contrast to LSI, it is based on a robust statistical foundation for an appropriate model of the data.

$$P(d, w) = \sum_{z \in Z} P(z) P(w|z)\, P(d|z) \tag{3}$$

LDA (Eq. 4) is a probabilistic model with three levels where each document of a collection is modeled as a finite mixture associated with multiple topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. Figure 1 (Blei et al. 2003) illustrates the graphical model representation of LDA where the outer plate expresses documents, while the inner plate represents the repeated choice of topics and words within a document.

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) \tag{4}$$

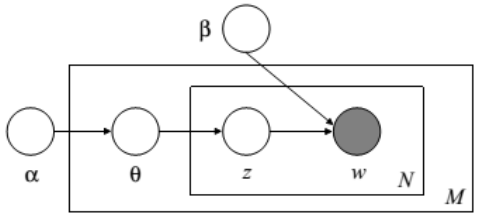

Figure 1. Graphical model representation of LDA (Blei et al., 2003).

3 THE METHOD

This section presents the proposed method of SARF (Semantic Analysis on Relevance Feedback) which comprises two main functions of concept extraction and concept manipulation. We first introduce the theoretical background of SARF, and then describe the development of SARF in detail.

3.1 The theoretical background of SARF

In the application of relevance feedback, conventionally, the information extracted from relevance feedback is utilized to build the user profile as the expanded query for document re-ranking. The term features of the feedback information, such as term frequency, document frequency, and term appearance, can be utilized for document and query representation. The feature of term appearance also can be utilized to identify the term relevance for irrelevant term elimination. In Wille's study (1992), he introduced a structure view of term relationships based on term appearance in a document set. In the structure, the term set forms the concept. This concept lattice structure can denote the two directions of concept relationships: the upper approximate concept (UAC) and the lower approximate concept (LAC). Figure 2 presents an example structure of concept lattice and the two directions of concept relationships for some documents.

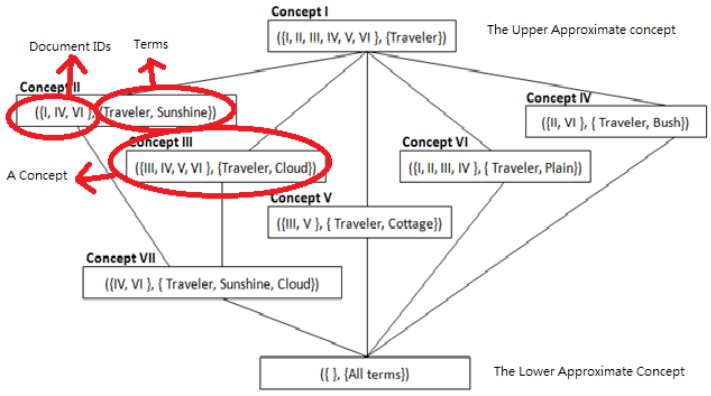

Figure 2. An example structure of concept lattice and the two directions of concept relationships for some documents

The structure of concept lattice could present useful information. For example, Figure 2 shows a set of documents about Europe journey formed by the concept lattice. The connection between the concept lattices reveals that Concept II and Concept III are the UACs of Concept VII. In contrast, Concept VII is the LAC against Concept II and Concept III. In the structure of the concept lattice as Figure 2 presents, the LAC plays a role of concept extension, and the UAC which discloses the partial existence of terms serves as the narrow sense of the concept. The partial existence of terms between the UAC and the LAC indicates the phenomena of concept propagation and concept fusion. Since the UAC only demonstrates the narrow sense and the LAC represents the general view of the concept for a specific object, both the UAC and the LAC individually could not offer the thorough information to satisfy the user's query needs about the object. Instead, the UAC, the concept propagation, and the concept fusion together could provide a broader and deeper sense of information about the concepts and might serve the user's query needs in a better way. In accordance with the inference as above, we propose a method, semantic analysis relevance feedback (SARF), which could analyze the semantic information concealed in the relevance feedback documents and apply the semantic information in the enhancement of document re-ranking.

3.2 The development of SARF

Figure 3 shows the flow of informational retrieval that has SARF in use. As the figure presents,

SARF consists of two main functions involving concept extraction and concept manipulation. In concept extraction, SARF performs two sub-functions involving concept construction and concept relevance identification. In concept construction, SARF deploys LDA (Eq. 5) to model the relevance feedback documents to construct the latent concepts. Then, the relevance of the concepts constructed by LDA is identified by the type of documents (relevant or irrelevant) in the relevance feedback. The relevant concept applies to candidate concept selection later.

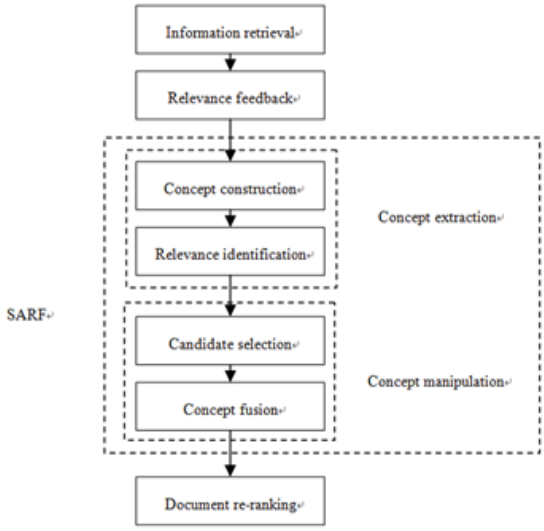

Figure 3. The Flow of SARF

$$T = \mathrm{LDA}(\alpha, \beta, \mathrm{TopicNum}) \tag{5}$$

Concept extraction will create many concepts that could be useful or not in terms of query representation. The function of concept manipulation of SARF has been designed to obtain the useful concepts. It comprises two sub-functions, including candidate concept selection and concept fusion. Candidate concept selection will filter out some relevant concepts identified earlier by the proportion of the term in the topic. The rest relevant concepts are selected as candidate concepts for concept fusion. At this stage, a candidate concept is used to represent a specific subject contained in the relevant documents, and all candidate concepts together indicate a user's query interest extracted from relevance feedback. Then, the sub-function of concepts fusion is exploited to produce each document's relevance to query interest with the concept hierarchy matter dealt. The major work is to calculate the similarity (Eq. 6) between a document and each candidate concept first, then, summarize all similarity values (Eq. 7) to obtain the document's score of concept fusion. The document's score of concept fusion, finally, is used for document re-ranking.

$$\mathrm{Sim}\left(T, \overrightarrow{d_i}\right) = \sum_{k=1}^{|T|} \frac{\overrightarrow{t_k} \times \overrightarrow{d_i}}{|\overrightarrow{t_k}| \times |\overrightarrow{d_i}|} = \sum_{k=1}^{|T|} \frac{\sum_{j=1}^{t} \left(\mathrm{Term}_{j,\overrightarrow{t_k}} \times \mathrm{Term}_{j,\overrightarrow{d_i}} \right)}{\sqrt{\sum_{j=1}^{t} \mathrm{Term}_{j,\overrightarrow{t_k}}{}^2} \times \sqrt{\sum_{j=1}^{t} \mathrm{Term}_{j,\overrightarrow{d_i}}{}^2}} \tag{6}$$

$$\text{the score of concept fusion of } d_i = \sum_{j=i}^{k} \mathrm{Sim}(T_j, \overrightarrow{d_i}) \tag{7}$$

The major difference between SARF and Rocchio's method is that SARF uses multiple relevant concepts to represent the diversity of relevance. Figure 4 is a further depiction to show that Rocchio's method utilizes the feedback documents to form a new query for relevance representation without the consideration of relevance diversity. In contrast, SARF fuses diverse candidate concepts to represent

the document's relevance for document re-ranking.

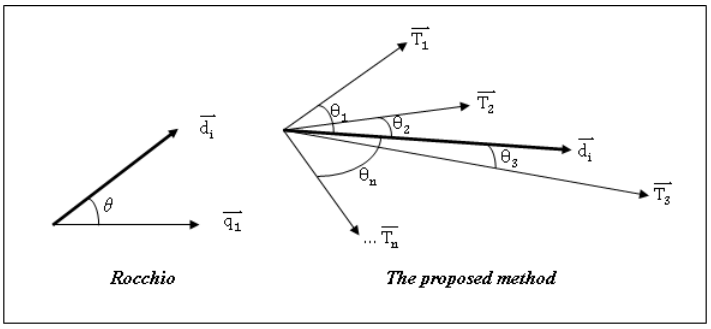

Figure 4. Comparison of SARF and Rocchio's method.

4 EXPERIMENTAL RESULTS

We have conducted some experiments in the evaluation of our proposed method of SARF, including (1) the optimal expanded terms for Rocchio's method, (2) LDA parameter selection for the optimal number of concepts, and (3) performance comparisons among SARF, vector model, and Rocchio's method. TREC CD4 & CD5 are selected as document collections. In the simulation of information retrieval, the title queries of topic 301 to 350, in TREC6, are utilized as the user's initial queries. The topic, with sufficient relevant documents (more than 100) in initial retrieval, is picked in our experiments for evaluation. The 15 topics which are appropriate for our experiment design are 301, 304, 306, 307, 311, 313, 318, 319, 321, 324, 331, 332, 343, 346 and 347.

Lemur is utilized as the information retrieval system in the experiments. It pre-processes the document collection and provides the basic term information, such as term frequency (TF) and document frequency (DF). The vector model and Rocchio's method are utilized as baselines. For performance evaluations, several measurements are utilized, including Precision (Eq. 8), Recall (Eq. 9), MAP (Eq. 10), and P@N (Eq. 11).

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrived iterms})} = \text{P}(\text{relevant}|\text{retrieved}) \quad (8)$$

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = \text{R}(\text{retrieved}|\text{relevant}) \quad (9)$$

$$\text{MAP}(\text{Q}) = \frac{1}{|\text{Q}|}\sum_{\text{j}=1}^{|\text{Q}|}\frac{1}{\text{m}_\text{j}}\sum_{\text{k}=1}^{\text{m}_\text{j}}\text{Precision}(\text{R}_{\text{jk}}) \quad (10)$$

$$\text{P@N} = \frac{\#(\text{relevant items retrieved top N documents})}{\#(\text{retrieved top N documents})} \quad (11)$$

In the first experiment, the number of expanded terms is examined to fetch the optimal performance of Rocchio's method in each selected topic. According to the conclusion in Salton & Backley (1997), the parameters of Rocchio's method are set as $\alpha = 1$, $\beta = 0.75$, and $\gamma = 0.25$. Then, the number of expanded terms of Rocchio's method is also examined and shown in Figure 5. Based on the results as Figure 5 has presented, QE_ALL is adopted in Rocchio's method for comparison with SARF.

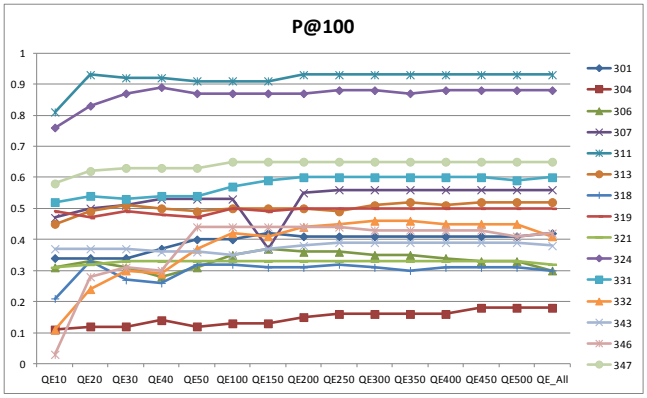

Figure 5. The trend of term number expended with Rocchio's method (P@100)

In the second experiment, we analyze the parameters (α and β) utilized in LDA to detect the optimal settings in our document collection for concept modelling in SARF. Two tests have been performed, including (1) the trend of topic distribution in different scaled α, and (2) the setting formulations utilized in Griffiths & Steyvers' study (2004). In the first test, the parameter α, enumerated within nine values (0.001, 0.01, 0.1, 1, 5, 10, 25, 50, 100), is exploited to reveal the trend of topic distribution, where the number of topic is from 1 to 20. The topic distribution (Dir(α)), from top left to down right, as shown in Figure 6, shows that the larger the value of α, the more uniform of the distribution. It is unsuitable for our experiment design. Therefore, we use another way to set the adequate values for parameters α and β.

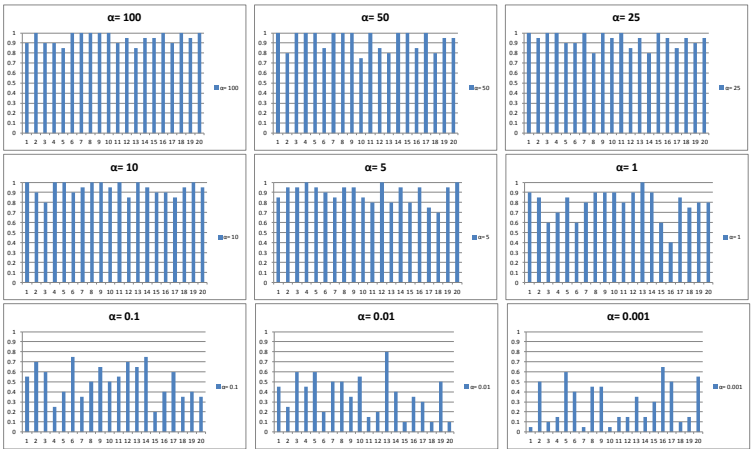

Figure 6. The topic distribution against the hyper-parameter α of LDA

Inspired by Griffiths & Steyvers' study (2004), the values of parameters α and β are formulated as Eq. 12, where T is the number of concepts. According to the observation in the consequences within the number of variable concepts, we set the number of concepts from 5 to 40 (5, 10, 15, 20, 25, 30, 35, 40). Table 1 presents the results in the different number of concepts and the optimal number of concept is 20.

$$\alpha = 50 / T, \qquad \beta = 0.01 \tag{12}$$

concepts	5	10	15	20	25	30	35	40
MAP	0.3676	0.393	0.3949	0.4003	0.3925	0.3968	0.3877	0.3937

Table 1. The results in different number of concepts

In third experiment, we deploy the parameter values fetched from the results of the first and the second experiments to compare the performance among SARF, vector model, and Rocchio's method with all expanded terms. Figure 7 shows the flow of the experiment. In simulating information retrieval, the

title query is used as the user's query for document retrieval in each topic. The retrieval results in each topic are viewed as the vector model results, and are utilized for relevance feedback. In the simulation of relevance feedback, the top 20 relevant documents and the top 20 irrelevant documents, in each topic, are picked as the user's feedback. These feedback documents are used for relevant information extraction by Rocchio's method and the proposed method, SARF. In Rocchio's method, the feedback documents are exploited in the modification of term weights to expand query terms to form a new query. The new query is then utilized in performing document re-ranking to re-sort relevant documents to higher ranks. In SARF, the feedback documents are processed by LDA to model the relationships among terms, concepts, and documents. For each topic, the top N concepts are fused by the mechanism of SARF to form a new query for document re-ranking. The re-ranking results performed by Rocchio's method and SARF separately, finally, are used for performance comparisons.

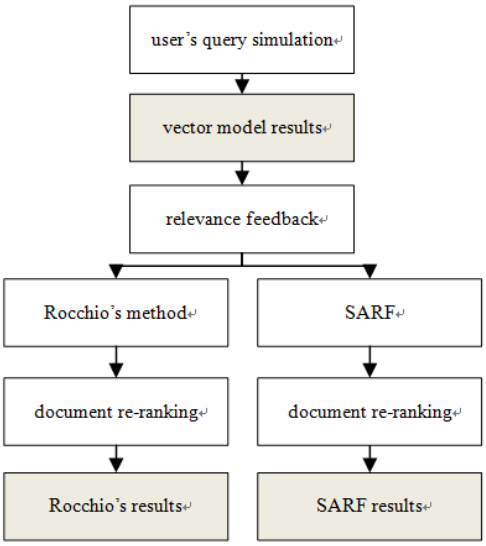

Figure 7. *The experiment flow for performance comparisons among SARF, vector model, and Rocchio's method*

Table 2 presents the original retrieval performance for and comparisons among SARF, vector model, and Rocchio's method. In the table, the column of imp_V presents the ratio of improvement for SARF comparing to the vector model, and the column of imp_R reveals the ratio of improvement for SARF comparing with Rocchio's method. Although Rocchio's method has performed fairly well in the carrying of relevant documents to higher ranks, SARF can make it even better.

	Vector Model	Rocchio	SARF	imp_V	imp_R
P@5	0.2933	0.68	0.7733	164%	14%
P@10	0.28	0.5933	0.66	136%	11%
P@15	0.2489	0.5644	0.6133	146%	9%
P@20	0.24	0.5567	0.5967	149%	7%
P@30	0.2267	0.5111	0.5778	155%	13%
top30_average	0.2577	0.5811	0.6442	150%	11%
P@100	0.2293	0.394	0.4593	100%	17%
P@200	0.212	0.3063	0.353	67%	15%
P@500	0.2084	0.2205	0.2457	18%	11%
P@1000	0.1873	0.1854	0.1903	2%	3%
top1000_average	0.2362	0.4457	0.4966	104%	11%
MAP	0.2343	0.3585	0.4003	71%	12%

Table 2. *Comparisons among SARF, vector model, and Rocchio's method on document re-ranking*

The comparison of precision-recall is shown in Figure 8, and it also denotes that SARF is more

impressive than both baselines.

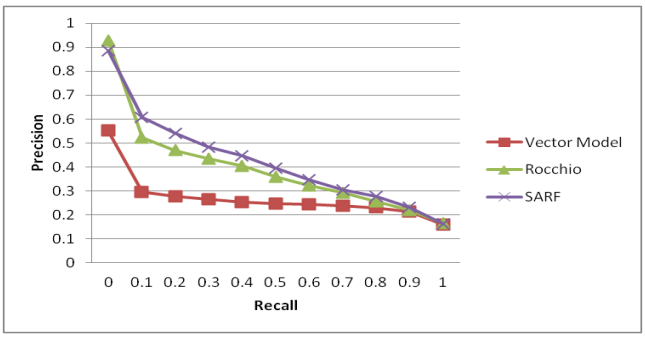

Figure 8. The comparison of precession-recall among SARF, vector model, and Rocchio's method

5 CONCLUSION

In this study, we proposed a method, SARF, to apply concept-based information residing in relevance feedback by way of semantic analysis. SARF fuses several relevant concepts, identified by the relevance with the initial query, to constitute the generic relevant concept, and adopt the generic relevant concept as new query for document re-ranking. The experimental results show that SARF has the capacity to identify more relevant documents from retrieved results than the baselines. The implication is that the semantic information residing in relevance feedback is worth of study and application. Therefore, this study has been conducting further experiments in the analysis of concept fusion.

Acknowledgements

This work was supported in part by the National Science Council, Taiwan, under the Grant No. NSC 102-2410-H-008 -041 -

References

Alshaar, R. (2008), 'Measuring the stability of query term collocations and using it in document ranking', University of Waterloo, Canada, 72 pages.

Amati, G. and Van Rijsbergen, C. (2002), 'Probabilistic models of information retrieval based on measuring the divergence from randomness', ACM Transactions on Information Systems (TOIS) 20(4), 357-389.

Balabanović, M. (1997), An adaptive web page recommendation service, in 'Proceedings of the first international conference on Autonomous agents', pp. 378-385.

Baziz, M.; Boughanem, M. and Aussenac-Gilles, N. (2005), IRIT at CLEF 2004: the english GIRT task'Multilingual Information Access for Text, Speech and Images', Springer, pp. 283-291.

Berners-Lee, T.; Hendler, J.; Lassila, O. and others (2001), 'The semantic web', Scientific american 284(5), 28-37.

Blei, D. M.; Ng, A. Y. and Jordan, M. I. (2003), 'Latent dirichlet allocation', the Journal of machine Learning research 3, 993-1022.

Carpineto, C.; De Mori, R.; Romano, G. and Bigi, B. (2001), 'An information-theoretic approach to automatic query expansion', ACM Transactions on Information Systems (TOIS) 19(1), 1-27.

Deerwester, S. C.; Dumais, S. T.; Landauer, T. K.; Furnas, G. W. and Harshman, R. A. (1990), 'Indexing by latent semantic analysis', JASIS 41(6), 391-407.

Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; Solan, Z.; Wolfman, G. and Ruppin, E. (2002), 'Placing Search in Context: The Concept Revisited', ACM Transactions on Information Systems 20(1), 116-131.

Griffiths, T. L. and Steyvers, M. (2004), 'Finding scientific topics', Proceedings of the National academy of Sciences of the United States of America 101(Suppl 1), 5228-5235.
Halliday, M. and Hasan, R. (1976), 'Cohesion in English, 1976', Longman, London.
Harman, D. (1992), Relevance feedback revisited, in 'Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval', 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 for Information Science 26(4), 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 26(5), 280-289.
Hofmann, T. (1999), Probabilistic latent semantic indexing, in 'Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval', pp. 50-57.
Koster, C. H. and Beney, J. G. (2007), On the importance of parameter tuning in text categorization'Perspectives of Systems Informatics', Springer, pp. 270-283.
Moldovan, D. I. and Mihalcea, R. (2000), 'Using wordnet and lexical operators to improve internet searches', IEEE Internet Computing 4(1), 34-43.
Moschitti, A. (2003), A study on optimal parameter tuning for Rocchio text classifier'Advances in Information Retrieval', Springer, pp. 420-435.
Nick, Z. Z. and Themis, P. (2001), 'Web search using a genetic algorithm', Internet Computing, IEEE 5(2), 18-26.
Okabe, M. and Yamada, S. (2005), 'Learning filtering rulesets for ranking refinement in relevance feedback', Knowledge-based systems 18(2), 117-124.
Robertson, S. E. and Jones, K. S. (1976), 'Relevance weighting of search terms', Journal of the American Society for Information science 27(3), 129-146.
Rocchio, J. J. (1971), 'Relevance feedback in information retrieval', .
Salton, G. (1971), 'The SMART retrieval system—experiments in automatic document processing', .
Salton, G. and Buckley, C. (1997), 'Improving retrieval performance by relevance feedback', Readings in information retrieval 24, 5.
Salton, G. and Lesk, M. E. (1968), 'Computer evaluation of indexing and text processing', Journal of the ACM (JACM) 15(1), 8-36.
Shanfeng, Z.; Xiaotie, D.; Kang, C. and Weimin, Z. (2001), Using online relevance feedback to build effective personalized metasearch engine, in 'Web Information Systems Engineering, 2001. Proceedings of the Second International Conference on', pp. 262-268.
Shek, E. C.; Vellaikal, A.; Dao, S. K. and Perry, B. (1998), Semantic agents for content-based discovery in distributed image libraries, in 'Content-Based Access of Image and Video Libraries, 1998. Proceedings. IEEE Workshop on', pp. 19-23.
Varelas, G.; Voutsakis, E.; Raftopoulou, P.; Petrakis, E. G. and Milios, E. E. (2005), Semantic similarity methods in wordNet and their application to information retrieval on the web, in 'Proceedings of the 7th annual ACM international workshop on Web information and data management', pp. 10-16.
Wille, R. (1992), 'Concept lattices and conceptual knowledge systems ', Computers & Mathematics with Applications 23(6ŸV9), 493-515.
Xu, J. and Croft, W. B. (1996), Query expansion using local and global document analysis, in 'Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval', pp. 4-11.
Yu, C.; Luk, W. and Cheung, T. (1976), 'A statistical model for relevance feedback in information retrieval', Journal of the ACM (JACM) 23(2), 273-286.
Zhang, B.; Du, Y.; Li, H. and Wang, Y. (2008), Query expansion based on topics, in 'Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on', pp. 610-614.