JAIDM
Journal of AI and Data Mining
Vol. 2, No .1, 2014, 79-84.

Query expansion based on relevance feedback and latent semantic analysis

M. Rahimi*, M. Zahedi

School of Computer and IT, Shahrood University of Technology, Shahrood, Semnan, Iran.

Received 15 May 2013; accepted 21 July 2013
**Corresponding author: Marziea.rahimi@shahroodut.ac.ir (M. Rahimi).*

Abstract
Web search engines are one of the most popular tools on the Internet, which are widely used by experienced and inexperienced users. Constructing an adequate query, which represents the best specification of users' information need to the search engine is an important concern of web users. Query expansion is a way to reduce this concern and increase user satisfaction. In this paper, a new method of query expansion is introduced. This method, which is a combination of relevant feedback and latent semantic analysis, finds the relative terms to the topics of user original query based on relevant documents selected by the user in relevant feedback step. The method is evaluated and compared with the Rocchio relevant feedback. The results indicate the capability of the method to better representation of user's information need and increasing significantly user satisfaction.

Keywords: *Query expansion, latent semantic analysis, relevant feedback.*

1. Introduction

The World Wide Web is a collection of vast amount of unlabeled, high dimensional and dynamic data. Web search engines as a quick and simple way to access to these data are widely used by experienced and inexperienced users. Most of the conventional search engines work based on text queries. Users enter some number of words as a query and search engine try to find documents related to the query. Therefore, success of the search process is greatly dependent on the user query; however, users enter broad or inadequate queries in most cases, which lead to retrieve irrelevant documents and user's dissatisfaction because of web search or intended information area inexperience. Users often enter topics of the intended information area and do not provide search engine with the other terms related to the topic.Query expansion is a method for reduction of this problem. Query expansion is the process of supplying the original user's queries with additional suitable terms. Many methods have been proposed for query expansion. These methods can be divided in two categories - global and local.

Thesaurus based query expansion [1, 2, 3] is a type of global methods. Each term on the query is expanded with its synonyms or related terms from the thesaurus. This type of methods because of expanding query by synonyms can significantly decrease precision especially in case of ambiguous query terms. Another problem of the thesaurus based methods is constructing and maintenance of a thesaurus. These processes are very expensive and time consuming. In case of such vast and dynamic spaces as the World Wide Web, these problems become more serious and effective. Based on these issues, thesaurus based methods are often used in special technology domains like medical technology.

Relevant feedback is a local method claimed to be most successful method between query reformulating methods. This method can be explained by steps below, which can go through one or more iterations.

- The user submits an initial query.
- Search engine returns an initial set of results.
- The user marks some of these results as relevant or optionally nonrelevant.

- The system produces a better specification of user information need based on the user relevance feedback.
- Search engine returns the revised results based on the new specification.

Some algorithms are proposed to implement relevant feedback in [4, 5]. These algorithms show that relevant feedback can significantly improve precision and recall. The method is based on this idea that despite the user inability to issue an adequate and suitable query for information need, it is easy for him/her to judge a particular document for relevant but in a web search engine, it is difficult to encourage people to judge resulted documents [6].

As mentioned earlier, query expansion is the process of supplying the original users' queries with additional suitable terms. There are two key elements for applying a query expansion method; resource from which the expanding terms is chosen and the algorithm employed for choosing these terms [7]. In this paper, resource is the set of relevant documents selected by the user in relevant feedback process and the algorithm is based on latent semantic analysis.

The proposed method tries to choose the suitable words from relevant documents selected by the user based on latent semantic analysis and then this method is evaluated on the Google search engine and is compared to the Rocchio relevant feedback.

1.1. Latent Semantic Analysis

Latent semantic analysis (LSA) is a mathematical technique for uncovering the underlying topic structure of documents. LSA is one of the most important methods for semantic retrieval [8]. It has been employed for the first time for information retrieval by Deerwester et al in 1990 [9] as an efficient technique to deal with polysemy and synonymy problem in information retrieval and quickly become popular. It is a widely used technique in knowledge discovery and representation, cognitive science, machine learning and many other areas. Some recent applications of LSA are topic detection [10], text summarization [11], FAQ retrieval [12] and text clustering.

Wei and Park use LSA along with genetic algorithm for text clustering. In this work, LSA is employed to reduce feature space so as to be appropriate for application of genetic algorithm [13]. Research [14] similarly uses LSA for dimensionality reduction. This work uses back-propagation neural networks for text categorization and for reducing the problem of slow training speed of this method, employs the LSA's capability of reducing dimensionality along with improving performance. Cohen et al employ LSA to construct a semantic space in which semantic associations between psychiatric terms are uncovered and these associations are used to segment and extract clinical concepts from psychiatric narrative [15]. Another recent paper introduces a method with combination of LSA and hidden Markov model for topic segmentation. In this paper, LSA is used for calculating similarity of a vocabulary term and a given topic term. These measurements then use as HMM emission probabilities [16].

LSA is not an artificial intelligence program; instead, it uses singular value decomposition (SVD) which is a matrix-analytical method to find base components space. Let X be a rectangular m by n matrix, SVD decomposes it into 3 matrixes as below:

$X = USV^T$

Where S is an n by n diagonal matrix containing singular values of X and U is an m by n matrix, whose columns are right singular vectors of X. V^T is an n by n matrix rows of which are left singular vectors of X. These singular vectors are taken geometrically as coordinates of the new n dimensional space. Singular values and their associated singular vectors are sorted in descending order of significance.

In information retrieval, X is term-document matrix, and each row of which is a term and each column is a document. Each entry of this matrix like x_{ij} indicates the weight of term i in document j. Weights are calculated based on TF-IDF.

2. Proposed method

According to many researches about web users' behavior, most users in the Web issue queries in the length of one term. The average of query length is 2 to 3 terms [17, 18]. These short queries can easily lead to conflict and inappropriate results. On the other hand, experienced users provide search engines with longer queries [19] and successful searches are performed with longer queries than unsuccessful searches. Query expansion with adding some appropriate terms to the initial user query can assist search engines to improve search performance and user satisfaction.

The proposed method consists of three steps: 1) User initial search, 2) User relevance feedback and 3) Query expansion. In the first step, user issue the initial query based on his/her knowledge and needed information. In the relevant feedback step, the user selects 5 relevant documents from the results of an initial search. As mentioned earlier, users often search topics of the domains of their

information need and ignore the other related terms of the domain. A relevant document collection can be used to extract other terms, which can be useful in producing more relevant results. The relevant document collection can be considered as a context for the original query terms. Each word in natural languages has several meanings and it is the context that determines which meaning of the word should be considered. A whole document is not a suitable choice for the context. Because a document consists of several topics and two words or terms are in the same context if they appear within a close neighborhood. Therefore, each document is divided into several windows in identical lengths. For preserving the sequence relationship of the words of the two successive windows, each window overlaps half of its next window.

The vector space model is applied to each window and a term-window matrix is constructed. LSA is applied to the matrix and $U \times S$ is calculated. Now, in the semantic space, the projected terms should be clustered. Each cluster is called topic cluster.

As mentioned earlier, latent semantic analysis uses SVD to decompose the term-document matrix X into three matrices U, S and V. Columns of U and V are singular vectors of X, which can be taken as dimensions of the semantic (topic) space. Each row of U associates with a term of vocabulary and $Y = U \times S$ is the projection of term by document matrix X in the topic space. Each dimension can be looked as a topic and each entry of matrix Y like y_{ij} is the importance of term i in topic j. With assigning each query to the topic, which has most influence in, some cluster of terms are constructed which specify each topic and are called as topic clusters.

The original user query is tokenized and stop words are removed to construct the set $q0$. Let C_j be a topic cluster and $q0_i$ be a member of set $q0$. And let EQ be an empty set.

$$q_{opt} = \{t_i \,|q0_i \in C_j \,\&\, t_i \in C_j\}$$

For each term $q0_i$ from vocabulary,t_i it is a member of vocabulary, and C_j is the topic cluster which $q0_i$ is assigned, and then all members of C_j are added to EQ. In some cases, the number of EQ's member may be large. In such cases, it is necessary to select some terms from EQ as extensions of original query. The global weights of terms is suitable for this aim i.e. m words with largest Global weights are selected to add to the original query. Global weight of each term is the number of selected relevant documents, which the term appears in.

2.1. Evaluation

The aim of the query expansion is increasing users' satisfaction. In many papers, precision and recall are used to evaluate the proposed information retrieval system. Precision is calculated at standard recall levels and then the precision-recall curve is interpolated. This procedure is performed for each test query and then the average is calculated over the test set of queries. Precision is the proportion of retrieved relevant documents over the whole retrieved documents and recall is the proportion of relevant retrieved documents over the whole relevant documents.

We cannot properly evaluate the users' satisfaction by this measure. One of the key elements of users' satisfaction is the rank of relevant documents in the resulted list. Users want to reach the best relevant document as soon as possible while browsing on the result pages is a boring and time consuming activity, which leads to dissatisfaction. Precision and recall do not consider this key element.

Another issue about the precision and recall is invisibility of not retrieved documents of the users to judgment. Some evaluation methods free of these problems has been introduced by researchers. One of these methods is binary preferences (Bpref) introduced by Beer and Moens in 2004 [20] as below:

$$Bpref = \frac{1}{R}\sum_{r} 1 - \frac{|N_r|}{R}$$

Where R is the number of relevant documents, N_r is a member of first R irrelevant documents ranked higher than the relevant document r.

For evaluating an information retrieval technique on the web, the number of result pages, which is considered, must be identified. The number of result pages viewed by the users is investigated in much research. Some research reveals that most users view only the first result page of the search engine. Jansen et al in an investigation on Excite search engine in 2000 found that 58 percent of users view only the first result page [21]. 19 percent view first two pages and 10 percent view the first three pages of result pages. Jansen and Spink in their research on search engine in 2003 appeared that 54 percent of users view first page, 19.3 percent first and second pages of result pages [22]. This percentage for research of Jansen et al which is performed on AltaVista search engine in 2005 are 72 percent for first page and 13 percent for first

two pages [23]. Based on the above research, most users view only first and second pages of result pages. Therefore, the evaluation of user satisfaction on a search engine is taken into account in this paper.

3. Experimental result

The proposed method has been evaluated on a test set of queries issued by participated users. Users are asked for issuing their initial queries and then judgment of resulted documents on the first two result pages as relevant or irrelevant. Bpref is calculated based on this judgment for each query.

Table 1. Example of proposed method's results

Original query	Expanded query by the proposed method	Bpref
Rose hybridization	Rose hybridization roses plants varieties cross process	0.5
Photography	Photography digital camera home articles techniques light	0.91
Owl wings	Owl wings feathers wing air sound prey	0.87

From the retrieved documents, five relevant documents are selected by the users in relevant feedback step. The LSA is applied on these relevant documents; expansion terms are selected and returned to the user. The user judges resulted documents of expanded query as the initial query results and Bpref is calculated for each new query. Results of the evaluation for some sample queries are reported in Table 1.

Rocchio algorithm [4] is the Standard relevant feedback algorithm. Refined query in this algorithm is calculated as below

$$q_m = \alpha q_0 + \beta \frac{1}{|D_r|} \sum_{d_j \in D_r} d_j - \gamma \frac{1}{|D_{nr}|} \sum_{d_j \in D_{nr}} d_j$$

Where α, β, γ are constant values attached to each term as its weight, d_j is a retrieved document, D_r is the set of selected relevant documents and D_{nr} is the set of selected irrelevant documents.

This algorithm can run more than one time but as reported in the literature, the most effective iteration is the first and other iterations do not improve the results significantly. This algorithm is implemented and has been applied on each test query. The user has judged results and Bperf has been calculated for each query. The results are reported in Table 2.

The average of Bpref on the test queries issued by the users is calculated for initial queries, Rocchio-resulted queries and produced queries of the proposed method. These values are displayed in Table 3 for comparison.

Table 2. Example of Rocchio method's results

Original query	Expanded query by the Rocchio method	Bpref
Rose hybridization	Rose hybridization roses plant seeds seed garden	0.64
Photography	Photography digital photos tips photo camera px	0.33
Owl wings	Owl wings Knoxville comments feathers owls published	0.75

Table 3. Mean Bpref for results of the considered methods

Method	Without query expansion	Query expansion by the proposed method	Query expansion by the Rocchio method
Bpref	0.43	0.76	0.59

Bpref value for the proposed method is the most. This method significantly improves user satisfaction. This improvement is dependent on improving both precision and ranking of relevant documents as shown in Table 4 for a sample query. In an equal condition with the proposed method i.e. equal number of selected relevant documents in relevant feedback step and equal number of expansion terms, cannot improve Bpref.

With more selected relevant documents, Rocchio algorithm may produce better results, but the relevant feedback step is time consuming and very boring for the impatient web users and selecting the larger number of documents intensifies the problem. Rocchio algorithm despite increment of precision and recall theoretically, does not provide any guarantee to higher rank the relevant documents than irrelevant ones. On the web, user satisfaction is highly dependent on showing more relevant documents first. Proposed terms by the Rocchio algorithm are more general terms and lead to a broader query while the proposed method

select more special terms and therefore a narrower query and better-ranked results. Difference between the proposed method and the Rocchio method results has tested for statistical significance by the paired t-test. The proposed method passed the test in the 95% level of significance.

The proposed algorithm does not necessarily select terms, which are more frequent. This algorithm selects the terms, which have similar frequency pattern to the user initial query terms in the selected relevant set of documents. For term selection, this method brings into account the number of relevant documents, which the word is appeared in too and as showed by the results, this method produce narrower queries and increases user satisfaction.

Table 4. An example of effect of the considered method on search results relevancy evaluated by the users

	Relevancy		
Method	Original	Rocchio	Proposed method
Query	Photography	Photography digital photos tips photo camera px	Photography digital camera home articles techniques light
Rank			
1	0	1	1
2	0	0	1
3	1	0	1
4	0	0	1
5	1	1	1
6	0	0	1
7	0	1	1
8	0	1	1
9	1	0	1
10	0	0	0
11	1	0	0
12	0	0	0
13	0	1	0
14	0	1	0
15	0	0	1
16	1	1	1
17	0	0	1
18	0	0	0
19	1	1	0
20	0	1	1
# Of relevant	6	9	12
Bpref	0.2	0.33	0.91

4. Conclusion

A new method of the query expansion has been introduced which analyzes frequency patterns of terms in the relevant documents selected by users from retrieved documents on an initial search based on users' original query. Based on this analysis, some topic clusters are constructed and the number of terms with the largest global weights is selected for adding to the original query. Resulted queries are evaluated based on Bpref measure which consider ranking of the retrieved documents along with the number of retrieved relevant documents for assessing users' satisfaction. The proposed method is compared to the Rocchio relevant feedback. This comparison shows that the proposed method produces queries more relevant to the user's information need and causes the lager number and higher ranking of the relevant documents. Therefore this method improves the user satisfaction significantly with little number of selected relevant documents.

References should be aligned with the journal style!!!

References

[1] Mandala R., Tokunaga T., Tanaka H., (2000). Query expansion using heterogeneous thesauri, Information Processing & Management, Volume 36, Issue 3, 1 May, Pages 361-378.

[2] Zazo A. F., Figuerola C. G., Alonso Berrocal J. L., Rodríguez E., (2005). Reformulation of queries using similarity thesauri, Information Processing & Management, Volume 41, Issue 5, September, Pages 1163-1173.

[3] Rocchio J., Relevance feedback in information retrieval. In: Salton G, ed. (1971). The Smart Retrieval System - Experiments in Automatic Document Processing", Prentice-Hall, Englewood Cli_s, NJ.pp. 313-323.

[4] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, (2009). An Introduction to Information Retrieval, Cambridge University Press.

[5] Efthimis E., (2010). Query Expansion. In Williams, Martha E. (Ed.), "Annual Review of Information Systems and Technologies (ARIST)", v31, pp 121-187.

[6] Valle-Lisboa J. C., Mizraji E., (2007). The uncovering of hidden structures by Latent Semantic Analysis", Information Sciences, Volume 177, Issue 19, 1 October, Pages 4122-4147

[7] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Beck, L. (1998). Improving information retrieval with latent semantic indexing. In Proceedings of the 51st annual meeting of the American society for information science, Vol. 25 (pp. 36–40).

[8] Masafumi Hamamoto, Hiroyuki Kitagawa, Jia-Yu Pan, Christos Faloutsos, (2005). A Comparative Study of Feature Vector-Based Topic Detection Schemes",Proceeding WIRI '05 Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration.

[9] Jen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang, and I-Heng Meng, (2005). Text summarization using a trainable summarizer and latent semantic analysis", Information Processing & Management Volume 41, Issue 1, January, Pages 75-95.

[10] Harksoo Kim, Hyunjung Lee, Jungyun Seo, (2007). A reliable FAQ retrieval system using a query log classification technique based on latent semantic analysis, Information Processing & Management Volume 43, Issue 2, March, Pages 420-430.

[11]- Wei Song, Soon Cheol Park, (2009). Genetic algorithm for text clustering based on latent semantic indexing", Computers & Mathematics with Applications, Volume 57, Issues 11-12, June, Pages 1901-1907.

[12] Bo Yu, Zong-ben Xu, Cheng-hua Li, (2008). Latent semantic analysis for text categorization using neural network", Knowledge-Based Systems, Volume 21, Issue 8, December, Pages 900-904.

[13] Trevor Cohen, Brett Blatter, Vimla Patel, (2008). Simulating expert clinical comprehension: Adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative, Journal of Biomedical Informatics, Volume 41, Issue 6, December, Pages 1070-1087.

[14] Filip Ginter, Hanna Suominen, Sampo Pyysalo, Tapio Salakoski, (2009). Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application", International Journal of Medical Informatics, Volume 78, Issue 12, December, Pages e1-e6.

[15] Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. (1999). Analysis of a very large web search engine query log. SIGIR Forum, 33(1), 6-12.

[16] Aula, A. (2003). Query Formulation in Web Information Search" Proc. IADIS WWW/Internet 2003, 403-410.

[18]- Buckley, C. and Voorhees, E. (2004). Retrieval evaluation with incomplete judgements", In Proceedings of SIGIR.

[19] Freund, L. & Toms, E.G. (2006). Enterprise search behavior of software engineers. Proc. SIGIR, 645-646

[20] Buckley, C. and Voorhees, E. (2004). Retrieval evaluation with incomplete judgements. In Proceedings of SIGIR.

[21] Bernard J. Jansen, Amanda Spink, Tefko Saracevic, (2000). Real life, real users, and real needs: a study and analysis of user queries on the web, Information Processing & Management Vol. 36, Issue 2, 1 March, Pages 207-227.

[22] Jansen BJ, Spink A. (2003). An Analysis of Web Documents Retrieved and Viewed. Fourth International Conference on Internet Computing; 2003; Las Vegas, Nevada. p. 65-9.

[23] Bernard J. Jansen, Amanda Spink, Jan Pedersen, (2005). A Temporal Comparison of AltaVista Web Searching, Journal of the American Society for Information Science & Technology, Vol. 56, No. 6, pp. 559-570.