International Journal of the Physical Sciences Vol. 7(24), pp. 3032-3036, 22 June, 2012
Available online at http://www.academicjournals.org/IJPS
DOI: 10.5897/IJPS11.1414
ISSN 1992 - 1950 ©2012 Academic Journals

Full Length Research Paper

Query expansion based on contextual meaning of the query terms

Marziea Rahimi*, and Morteza Zahedi

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

Accepted 23 April, 2012

In this paper, a new method for query expansion is proposed which expands the original query based on the contextual meaning of it. User provides the method with an original query and a set of relevant documents. The term space extracted from the relevant documents by the vector space model is transformed to a topic space by means of latent semantic analysis. The terms' projections in the topic space are grouped to some clusters called topic clusters. Expansion terms are the terms which have the same topic as the query terms. The results are evaluated by users in Google search engines and evaluation results show that the proposed method improves significantly the user satisfaction.

Key words: Query expansion, term clustering, contextual topic, latent semantic analysis, cluster analysis.

INTRODUCTION

Nowadays, world wide web is one of the most important resources of information in human societies. Web search engines are simple tools for access to this dynamic and huge information resource. Current conventional search engines work based on user queries which are a set of word or terms. Therefore, the construction of an adequate query which represents the best specification of their information need is one of the biggest concerns of web users.

As reported in many studies (Jansen et al., 2005; Jansen et al., 2000; Jansen and Spink, 2003), in about 40% of web searches, query length is one term and the mean is 2 to 3 words, therefore, users' queries on the web are too short and because of the ambiguity of words in natural languages, a short query leads to conflict and imprecise results. Inexperience in two domains is the reason of users' short and inadequate queries; inexperience in web search and domain of needed information (Holscher and Strube, 2000; White et al., 2009).

Supporting the users' original queries by some relevant specific terms can lead to the construction of a better query which is a better representation of user information need, and therefore, more precise results. This process is called query expansion. Query expansion methods are divided into two branches; based on search results and knowledge structures (Efthimiadis, 1996) as shown in the Figure 1. The first branch consists of relevance feedback methods that claimed to be the most effective category of query expansion methods (Manning et al., 2009) and the second is related to thesaurus-based methods (Hartmann and Brown 2005; Mandala et al., 2000; Bechhofer and Goble, 2001). Thesaurus-based methods are expensive and time-consuming methods in construction and maintenance, and therefore, are more suitable for specific domains like medicine.

Relevance feedback implementations claimed to be the most effective methods of query expansion. This technique is employed in this paper to constructing a set of relevant documents. The set of relevant documents is the resource for selecting expansion terms. Latent semantic indexing (LSA) and term clustering are employed for this aim. Most of the relevance feedback methods select expansion terms only based on their frequency in the relevant document set and the meaning of words is not directly brought into account. In this paper, a method is proposed which cluster the terms based on their contextual meaning in the relevant document set and then select the expansion terms from the clusters which are relevant to the original user query based on terms meaning.

*Corresponding author. E-mail: mr_ir26@yahoo.com or mrahimi26@shahroodut.ac.ir.

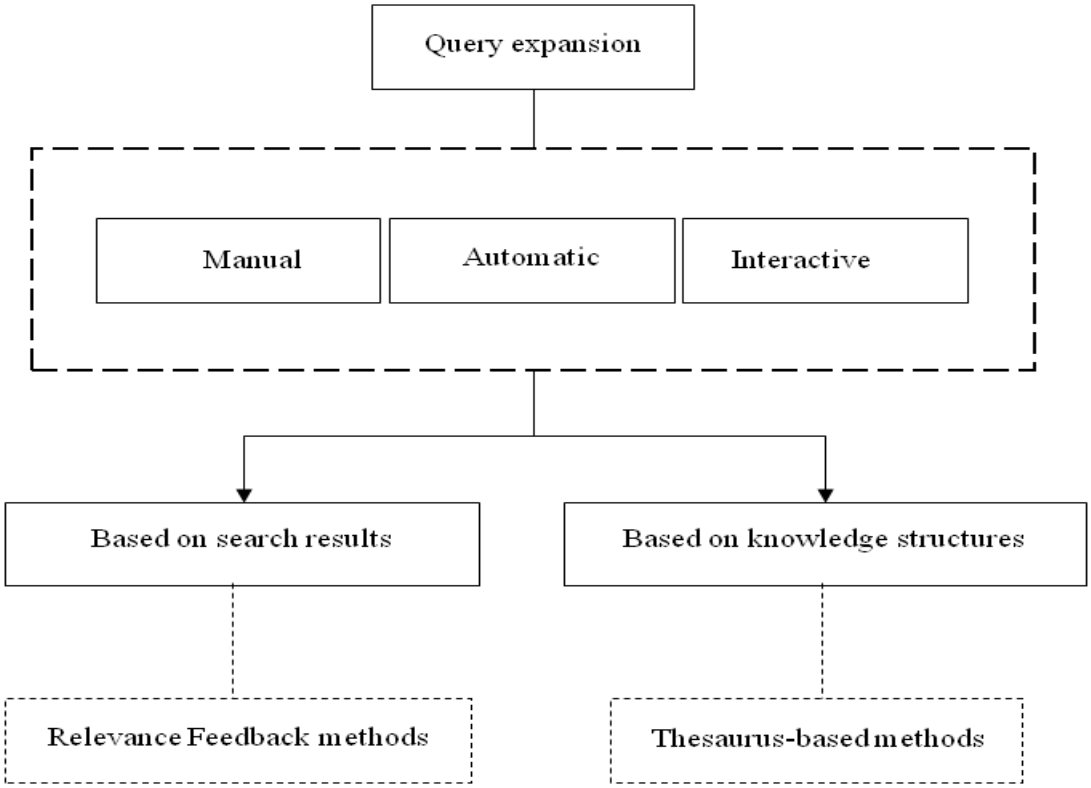

Figure 1. Classification of query expansion methods.

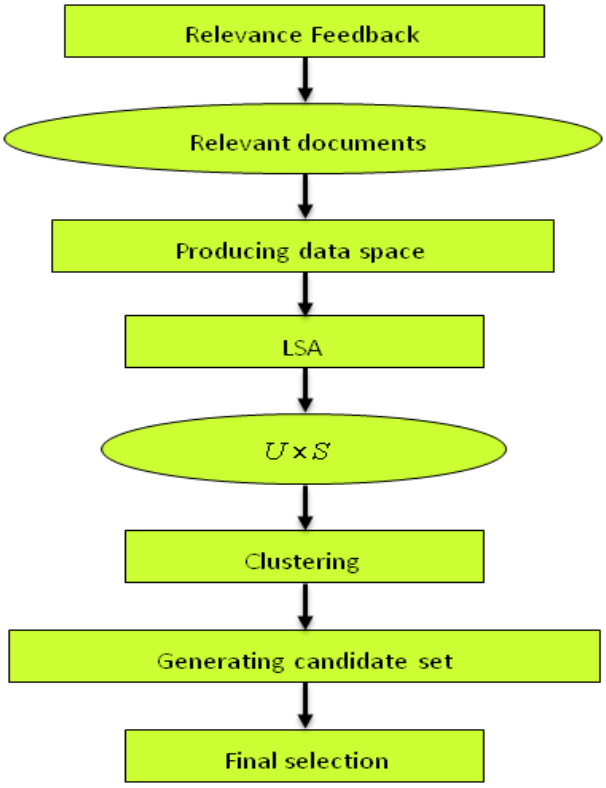

Figure 2. The diagram of the proposed method.

PROPOSED METHOD

As shown in the Figure 2, this method includes a relevance feedback step in which, the user select 5 relevant documents from the results of an initial search. 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 retrieving relevant documents. Dividing the relevant document collection into several domains is the solution. Term clustering (Gliozzo and Strapparava, 2009; Bassiou and Kotropoulos, 2011) technique can be used in this operation.

Clustering is an unsupervised method for grouping some objects in a way that objects in the same cluster are similar to each other in an aspect and members of different clusters are different based on the same aspect. Clustering algorithms can be applied to a wide variety of objects, such as terms. Grouping terms based on their frequency patterns in a text collection is the term clustering.

A topic cluster is a cluster whose members (terms) are related to an identical topic. In natural languages, each term has several meanings and the precise meaning of a term is dependent on the context and can specify by extracting other related terms which have related meanings to the terms, based on the context. A topic cluster is a suitable group of these related terms which can specify the meaning of a query term for a search engine and lead the search engine to the correct topic and documents which are precisely relevant to the query terms. For this aim, vector space model is a suitable model for representation of words.

Latent semantic analysis is a technique for uncovering the underlying meaning relationships between terms and documents. LSA is firstly introduced by Deerwester et al. (1998), for indexing in information retrieval. In this application, the technique is called

$$\begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} u_{1,1} \\ \vdots \\ u_{m,1} \end{bmatrix} \dots \begin{bmatrix} u_{1,k} \\ \vdots \\ u_{m,k} \end{bmatrix} \end{bmatrix} \times \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_k \end{bmatrix} \times \begin{bmatrix} [v_{1,1} & \dots & v_{1,n}] \\ & \vdots & \\ [v_{k,1} & \dots & v_{k,n}] \end{bmatrix}$$

Figure 3. Singular value decomposition.

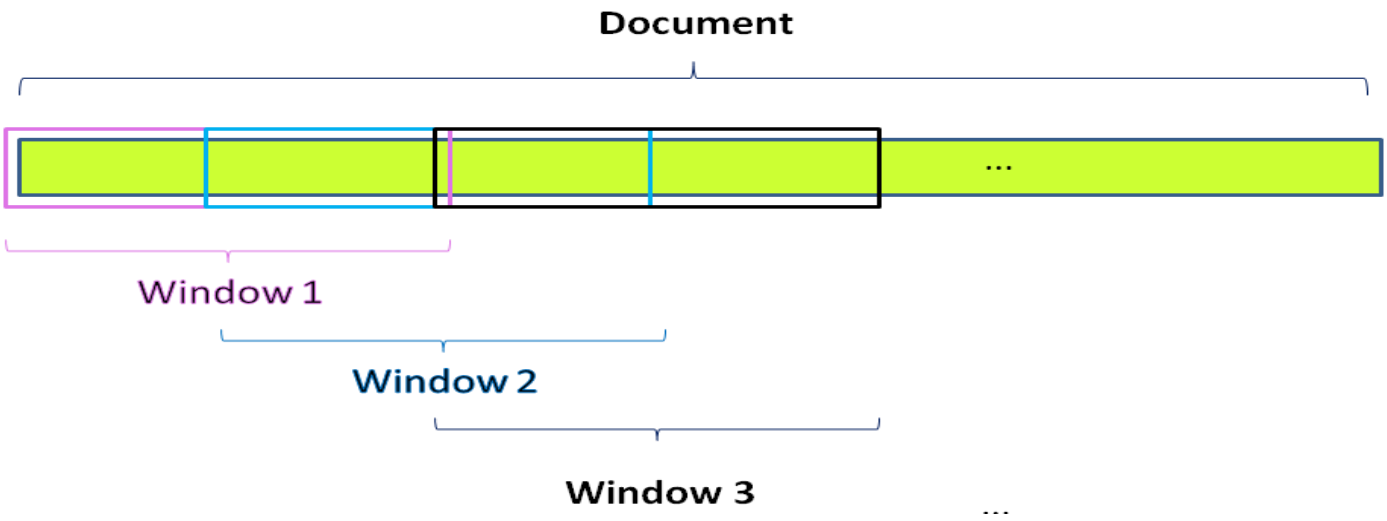

Figure 4. Each document is divided into several overlapped and identical windows.

$$\mathbf{t}_i^T \rightarrow \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} \quad \mathbf{d}_j \downarrow$$

Figure 5. Term-window matrix.

latent semantic indexing. This technique is widely used in many applications of information retrieval, like topic detection (Hamamoto et al., 2005; Gintera et al., 2009; Kuhna et al., 2007), text summarization (Yeh et al., 2005) and text and document clustering (Song and Park, 2009; Weia et al., 2008).

LSA uses singular value decomposition which is a matrix analysis technique. Singular value decomposition (SVD) decomposes the input matrix into 3 matrices U, S and V^T as shown in Figure 3. Columns of U and V are singular vectors of X and main diagonal of S contains associated singular values of these vectors. U and V are orthonormal and therefore each column of these two matrices can be considered as a dimension of a new space.

Each row of X indicates a term and each column of which indicates a document. Words correlation can be calculated by XX^T which is equal to US^2U^T as shown in Equation 1. Therefore, US is associated to each word in the new space which is called “topic space”. Each dimension in the new space can be considered as a topic and US is a term in the topic space.

$$XX^T = USV^TVSU^T = US^2U^T \quad (1)$$

Each clustering technique can be used to cluster the terms’ projection in the LSA or topic space.

K-means and hierarchical clustering are two conventional methods in text and term clustering which are used in this paper for term clustering in the topic space. For the two methods, the cosine similarity is used to calculate the similarity of each pair of term vectors.

The relevant document collection can be considered as a context for the original query terms. As mentioned earlier, each word in natural languages has several meanings and it is the context that determines which meaning of the word should be considered. Is a document, a suitable choice for the context? No; because a document consists of several topics and two words or terms that appear in the same context if they appear in a near neighborhood. Therefore, each document is divided into several windows in an identical length. For preserving the sequence relationship of the words of the two successive windows, each window is overlapped by half of its next window as shown in Figure 4.

Table 1. Some example expanded queries.

	Original	Rocchio	LSA + Hierarchical	LSA + K-means
Query	Apartment plants	Apartment Plants Garden Plant Home Gardening Care	Apartment Plants Home Garden gardening Shop Inside Comments	Apartment Plants Plant Care Light Water Grow Easy
Bpref	0.20	0.55	0.15	0.62
Query	Digital photo	Digital Photo Photography Photos Tips Camera	Digital Photo Tips Photographers Photographer camera People	Digital photo Photography Camera photographers photos
Bpref	0.26	0.35	0.78	0.51
Query	Conference Information Science	Conference Information Science International Computational End	Conference Information Science Papers contact Home	Conference Information Science Papers Research contact International
Bpref	0.31	0.48	0.08	0.57

The vector space model is applied to each window and a term-window matrix is constructed as Figure 5. LSA is applied to the matrix and US is calculated. Now, in the semantic space, the projected terms should be clustered. Each cluster is called topic cluster.

According to our experiments, the best results are produced by setting the number of clusters to 4.After construction of topic clusters, the candidate terms should be selected. The candidate terms are terms which are assigned to the same cluster which the original query terms are assigned. For selecting terms from the set of candidate terms, the global weight is adopted. Global weight of each term is the number of windows which the term has occurred in.

EVALUATION

In this paper, binary preferences (Buckley and Voorhees, 2004) which is a rank-based method and can deal well with incomplete judgments is employed to evaluate the proposed method. The equation of this method is modified for leading to more understandable resulted values. It is calculated by Equation 2. Where R is the number of relevant documents, N is the number of irrelevant documents and N_r is a member of first R irrelevant documents ranked higher than the relevant document r. High Bpref means that an algorithm returned more relevant results in higher ranks.

$$\mathbf{Bpref} = \frac{1}{\mathbf{R}+\mathbf{N}}\sum_{\mathbf{r}} \mathbf{1} - \frac{|\mathbf{N_r}|}{\mathbf{R}} \quad (2)$$

EXPERIMENTAL RESULTS

Based on some information needs, users formulate their queries and then submit their issued queries on Google search engine and from the results, select 5 documents, which have more relevancy to the users' information need in their opinion, as the relevant documents. The term-window matrix and then the topic clusters are constructed and the expansion terms are selected. Expanded query is submitted to the Google search engine and the results of the queries in the first result page are judged by the users and evaluated by binary preferences method.

Table 1 shows the results for the original query, expanded query based on the K-means clustering and expanded query based on the hierarchical clustering method for three example queries. This table includes the result of the Rocchio relevance feedback method (Salton and Buckley, 1990) for comparison. This method works based on Equation 3, where $\alpha,\ \beta$ and γ values are constant values attached to the equation terms as their weights, 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.

$$q_m = \alpha q_0 + \beta \frac{1}{|D_r|}\sum_{d_j \epsilon D_r} d_j - \gamma \frac{1}{|D_{nr}|}\sum_{d_j \epsilon D_{nr}} d_j \quad (3)$$

Table 2. Comparison of the methods.

Method	Original	Rocchio	LSA + k-means	LSA + hierarchical clustering
Mean of Bpref on the test set	0.37	0.46	0.53	0.50

As shown in Table 1, the expanded queries produced based on the K-means clustering method includes more precise and relevant terms and the value of Bpref for this method is larger than the others. For example, the expansion terms of the query “Apartment plants” by the Rocchio method are “apartment plants garden plant home gardening care”, which are very public. In contrast, these terms for the K-means are “plant care light water grow easy” which is precisely relevant to “Apartment plants”. Improvement of results is obvious in the table.

Difference of the results of the four states is tested for statistical significance by the paired t-test. The proposed method when using the K-means clustering method, pass the test in the 95% level of significance. But the P-value for the hierarchical clustering is greater than 10%, and therefore, it is not statistically significance. Table 2 shows the mean values of Bpref for comparison. These values are calculated on a test set of 141 queries which are constructed by users and results of each query are judged by the user who issued the query.

Conclusion

In this paper, a new method of query expansion is introduced which expand an original query based on contextual meaning of it. User issued some terms as the query and indicates his/her mean of the terms with providing search engine with some relevant documents. The term space extracted from the relevant documents is transformed into a topic space by the LSA technique. In the topic space term, clustering is performed by the hierarchical and K-means clustering methods. The expansion terms are selected from topic clusters which is containing the original query terms. Therefore, the proposed method expands the original queries with some terms which are relevant to an identical topic with the user’s original query terms and this relevancy is extracted from the context defined by the set of user-selected relevant documents.

The proposed method by both clustering methods, increments the user satisfaction which is evaluated by Bpref method based on the user judgment of results. The results of the proposed methods are compared with the Rocchio method and the initial query results. The proposed method which uses the K-means clustering method produces statistically significant improvements in the initial and Rocchio results.

REFERENCES

Bassiou N, Kotropoulos C (2011). Long distance bigram models applied to word clustering. Pattern Recognit., 44(1): 145-158.

Bechhofer S, Goble C (2001). Thesaurus construction through knowledge representation. Data & Knowledge Eng., 37(1): 25-45.

Buckley C, Voorhees E (2004). Retrieval evaluation with incomplete judgments. Proc., SIGIR.

Deerwester S, Dumais ST, Landauer TK, Furnas GW, Beck L (1998). Improving information retrieval with latent semantic indexing. Proc. 51st Annual Meeting. Am. Soc. Inf. Sci., 25: 36-40.

Efthimiadis EN (1996). Query Expansion. In Williams, Martha E. (Ed.), Annu. Rev. Inform. Syst.and Technol., (ARIST), 31: 121-187.

Gintera F, Suominena H, Pyysalob S, Salakoskia T (2009). Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application. Int. J. Med. Inform., 78(12): e1-e6.

Gliozzo A, Strapparava C (2009). Semantic Domains in Computational Linguistics. 1st ed Springer-Verlag Berlin Heidelberg, pp 41-43

Hamamoto M, Kitagawa H, Pan JY, Faloutsos C (2005). A Comparative Study of Feature Vector-Based Topic Detection Schemes. Proceeding WIRI '05 Proc. Int. Workshop on Challenges in Web Information Retrieval and Integration.

Hartmann RRK, Brown K (2005). Thesauruses. Encyclopedia of Language & Linguistics.2nd ed. Elsevier. Oxford, pp. 668-676.

Holscher C, Strube G (2000). Web search behavior of Internet experts and newbies. Computer Networks, 33: 337-346.

Jansen BJ, Spink A (2003). An Analysis of Web Documents Retrieved and Viewed.4th Int. Conference on Internet Computing; Las Vegas, Nevada, pp. 65-69.

Jansen BJ, Spink A, Pedersen J (2005). A Temporal Comparison of AltaVista Web Searching. J. American Society Inform. Sci. Technol., 56(6): 559-570.

Jansen BJ, Spink A, Saracevic T (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. Inform. Process Manag., 36 (2): 207-227.

Kuhna A, Ducasseb S, Gîrba T (2007). Semantic clustering: Identifying topics in source code. Inf. Software Technol., 49(3): 230-243.

Mandala R, Tokunaga T, Tanaka H (2000). Query expansion using heterogeneous thesauri. Inf. Process Manag., 36(3): 361-378.

Manning CD, Raghavan P, Schütze H (2009). An Introduction to Information Retrieval. Cambridge University Press, pp 177-194

Salton G, Buckley C (1990).Improving Retrieval Performance by Relevance Feedback. J. Am. Soc. Inf. Sci., 41(4): 288-97.

Song W, Park SC (2009). Genetic algorithm for text clustering based on latent semantic indexing. Comput. Math. Appl., 27(11-12): 1901-1907.

Weia CP, Yangb CC, Lin CM (2008). A Latent Semantic Indexing-based approach to multilingual document clustering. Decis. Support Syst., 45(3): 606-620.

White RW, Dumais ST, Teevan J (2009). Characterizing the Influence of Domain Expertise on Web Search Behavior.Proc. Second ACM Int. Conference on Web Search and Data Mining (WSDM '09). Barcelona. Spain.

Yeh JY, Ke HR, Yang WP, Meng IH (2005).Text summarization using a trainable summarizer and latent semantic analysis. Inform. Processing Manag., 41(1): 75-95.