International Journal of Psychosocial Rehabilitation, Vol. 25, Issue 03, 2021
ISSN: 1475-7192

USE OF INVERSE-TERM FREQUENCY (ITF) AND RELEVANCE FEEDBACK TO IMPROVE QUERY EXPANSION

Pawanjit Kaur
Guru Kashi University, Talwandi Sabo

ABSTRACT

The field which is full of concerned with the structure, analysis, institution, space and searching is Retrieval of information. It has now become an essential field of investigation and research under computer science because of the amount of data available in full text, hypertext, administrative text, directory, numeric, or bibliographic text has increased dramatically. There are several points of information or data retrieval system on which it is compulsory to conduct a proper research work. The objective of this research is to investigate the query expansion procedure using inverse-term frequency to improve the efficiency and accuracy of the information retrieval system. As the method of evaluation of query expansion, we will remove unrelated, redundant and ambiguous words from the retrieved document based on user- query. In proposed work, we introduce a new method for query expansion (QE) which is based on inverse term frequency with relevance feedback. Fetching the top revive documents use as in relevance feedback for additional QE terms and constructing candidate terms. Process of scoring method assigns score to unique terms and applying inverse term frequency (itf) to produce the rank list of terms. These terms will filter through semantic action and reweighting produce updated (expanded) query which will again send to search tool.

Keywords-Inverse-term frequency, Query Expansion, Precision, KLD-mean, Sementic similarity, term-pooling

I. Introduction

Information retrieval is the concept of finding unstructured material using any user's query based on required information, from a perfect collection of documents. In reality, none of the data are really "unstructured". If you count the hidden linguistic structure of natural

languages, this is true for all text data. However, even if the desired concept of structure is overt structure, most of the text has been structure, such as headings, paragraphs and footnotes, which is represented in documents by precise mark-up. IR is also used to ease of "semi-structured" search such as getting a document whose the title contains Java and the body contains threading.

II. Objectives of Information Retrieval Systems

Main objective of an IR System is to reduce the overhead of user desired information. User spends time can be overcome in all steps important to read essential required information. Information can be useful for users only under certain conditions.

i. To reduce the information overload from the system, by providing precise and exact 'user query based' information.
ii. Finding best available relevant information resources based on user formulated query, to reduce overall time-spent of the user for finding what they need.
iii. Based on syntactic and semantics structure of the query, search results may be different for the exact same keywords. This will provide more area for expansion of query to get most prominent results.
iv. IR systems try to overcome the problem of arranging and searching a large corpus of data presented to them from various sources.

III. Performance Evaluation of Information Retrieval System

The two main desired properties of measurement of search effectiveness are Recall (the proportion of relevant documents that are retrieved) and Precision (the proportion of retrieved document that are relevant). Another entity (which is derived from both precision and recall) can also used as an evaluation parameter and that entity is called F-Measure. There are three major evaluating factor of performance measurement of IR system.

IV. Related work

F. B. D. Paskalis et al. published a paper, "Word Sense Disambiguation in Information Retrieval Using Query Expansion". In this framework they presented a novel framework for QE to handle ambiguous queries and shows that exploratory queries are very difficult to judge. They proposed an Extended WordNet based WSD algorithm for disambiguaty of different part of speech category.

M. Song et al. presented a paper on “Integration of association rules and ontologies for semantic query expansion” in this paper author proposed the semantically based QE method which merges association rules with NLP techniques. They used the fusion QE based algorithm and merged to association rules using NLP techniques. Result shows that the algorithm is useful for meanings as well as language acquiring based properties of dynamic corpus.

Y. SONG et al. published a paper, “Simple Weighting Techniques for Query Expansion in Biomedical Document Retrieval” in this paper author anticipated two methods to pick up the performance and effectiveness in medical domain retrieval. Query having short bio word are expanded with its synonyms and multiword terms. Conventional method a document ranks highly contain because the expression has additional change to be matched with a query and provide unsatisfactory result. The method of stabilize the weights for query terms in an extended word in biomedical vocabulary. The biasing term using itf also produce to be successfully stats. This was improving the biasing power and its application performed on MEDLINE.

A. Ali-Abdelmgeid published an article “Query Expansion Technique to Improve Document Retrieval”. The major objective to use QE based technique to get an appropriate added term that develops a new efficient query for retrieval effectiveness. For QE they used two types of approaches, one is thesaurus based similarity expansion and other is similarity of local feedback. Thesaurus based similarity calculates the relevance between terms and queries to exchange the tag of documents and terms in recovery. While the local feedback method modified queries of the initial recovery.

N. Bansal et al. published a paper on "Measure-driven Keyword-Query Expansion". In this article he proposed new searching model, which has been getting an outburst in available data. In this circumstance, it commonly expressed either by unambiguously or totally. In this paper they proposed a novel searching method, which allowed the keyword based result. Their query-driven and domain-neutral approach utilized for co-occurrence where user ratings in order to find significant top-k query expansions of the original result.

D. Bernhard was published a paper on "Query Expansion based on Pseudo Relevance Feedback from Definition Clusters", he proposed a innovative method for QE based on PRF. They focused on the static and dynamic dictionary definition for QE. The definition clusters are made across different English lexicon resources. The method as followed: (i)

to build definition of clusters, (ii) To compare different expansion methods by using feedback. The modification in the query provided by feedback from definition of clusters shown significant improvement of the retrieval results.

H. J. Peat et al. was published a paper on " The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems ".In this paper they proposed a method to recognize the ranking terms that are similar to specified user query:., author has recognized a subset of term co-occurrence data as a basis for a AQE. Since query-term tend to have high frequencies given improved results than queries which had been elongated by any other methods.

J. Singh and A. Sharan have analysis on various relevance methods then he observed that every separate extension terms selection method which had some flaws and potency. To resolve these findings, they published a paper on "Relevance Feedback Based Query Expansion Model Using Borda Count and Semantic Similarity Approach", In this paper, to increase the performance by entity based QE terms – selection procedure, Borda count rank aggregation approach was applied for merging various terms selection methods and with the help of semantic similarity to select duplicate terms. The major approach was used to combine QE based terms selection methods to enhance overall performance of IR system.

V. Problem formulation and proposed work

We focus on a new method for boosting ranking factor which can be used to improve indexing of retrieved documents. Assumption is that top ranking documents in the first time relevant, after that feedback strategy uses in which weighting can be find through well known method inverse term frequency (itf) and various scoring methods like KLD.

VI. Dataset

This is an important technique of Information Retrieval [24]. Its purpose is to improve the relevancy and quantity of the retrieved results. In this work, term- biased types queries will be constructed from the available corpus named is tcorpus or we will develop a new corpus if needed. tcorpus is the text based database which is collection of encyclopedia of various dictionaries of English language. In which we can search query as input in English terms using BM-25 tool. An efficient BM25 has been used in similarity function for gathering a set of retrieval of documents.

VII. Procedure of Proposed Framework

Our framework can be categorized into two categories: initially we fired query over existing framework, thereafter we have been processed for obtaining results into itf based module as proposed framework.

- Apply Okapi- With respect to a query of customer, BM25 similarity function is essential for retrieving the appropriate document
- All the unique and relevant terms of top 'n' retrieved documents attained from step are chosen to create a term pool.
- Now, KLD score procedure is applied to give unique scores upon terms from term-pool and these terms are known as candidate (associated) terms.
- After getting the candidate terms, Inverse term frequency of these candidate terms is calculated, and a ranked list of the final terms is formulated. This ranked list is sorted according to the decreasing relevancy (higher the ITF score, more will be the term is relevant), as one go from top to bottom. To widen the user inquiry, some of the top ‘n' terms from this final candidate term pool are used. This method is known as ITF-based query expansion.
- To avoid the problem of query drifting, semantic connection approach is mainly used to sort out all inappropriate semantic terms from the list of candidate terms set gained from ITFBQE approach. After using semantic analysis, ITF with semantic-based procedure is called ITFSBQE

VIII. Summary of Experimental Results

The overall study of the project on Inverse Term Frequency based Query Expansion and its semantic based counterpart is summarized in the following points:

i. The individual ranking method KLD based query expansion performs better than other available approaches, hence it’s the most beneficial for our approach. OKAPI-BM25 is also used as a matching factor for the input query.

ii. ITF based query expansion perform better than the available methods especially KLD based query expansion and Borda based query expansion (BBQE) in the term of average precision on both FIRE and TREC datasets.

iii. ITF based query expansion method, when combined with semantic filtering, and then ITFSBQE outperform the all available methods in terms of average precision on both tcorpus datasets.

iv. The overall results shows that the proposed ITF based query expansion can provide highly precise documents to the user based on the given query to information retrieval tool. If the method is used along with semantic filtering then it gain a significant amount of performance boost compared to ITF based query expansion and other methods.

IX. Future work

Probing deeper, results in this thesis also connect a foundation for future scope in the area of information retrieval using query expansion. Finally, we have provided a detail of future scope which will provide an assistance to improve results of query expansion procedure. In this proposed framework we will implement the effective query expansion method to enhance performance of IR system by using the various selection procedures in real data set and to calculate average precision values getting from data set like and tcorpus based . Further, robustness of the proposed method can be check by using many other real databases like other languages database or in scientific database and more expansion terms. The average precision can also be more accurately calculated for higher number of the retrieved documents.

X. Conclusion

Our approach is limited to the average precision of the system, while other performance evaluation factors such as recall and F-Measure can also be calculated to provide better prove of the correctness of the system. A statistical analysis can also be done to prove that the improvements are statistical significant. The ITF based query expansion can be used in place of other term biasing methods, once it’s error free. The semantics similarity approach can also be further refine by using other available semantics matching schemes and evaluating their performance with ITF based query expansion.

XI. References

[1] C. Mooers, “Information retrieval viewed as temporal signaling”, in the Proc. of the International Congress of Mathematicians, Vol 1, pp 572- 573, 1950.

[2] D. Lauren and B. Joseph, “Information Retrieval and Processing, Melville”, SIGIR, Vol. 11, No. 2, pp 07 – 09, Fall 1976.

[3] H. J. Peat and P. Willett, “The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems”, ASIS, Vol. 42, No. 5, pp.378–383, June 1991.

[4] Z. Wu and M. Palmer, “Verb semantics and lexical selection,” in Proceedings of the Annual Meeting of the Associations for Computational Linguistics, pp. 133–138, 1994.

[5] P. Resnik, “Using information content to evaluate semantic similarity in a taxonomy,” in Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI ’95), vol. 1, pp. 448–453, Montreal, Canada, 1995.

[6] C. Leacock and M. Chodorow, “Combining local context and WordNet similarity for word sense identification,” in WordNet. An Electronic Lexical Database, pp. 265–283, MIT Press, Cam- bridge- Mass, USA, 1998.

[7] J. Ricardo Baeza-Yates and B. Ribeiro-Neto, "Morden Information Retrival" in ACM Press, Addison Wesley Longman Ltd.Engaland, 1999.

[8] A. M. Lam-Adesina and G.J.F. Jones, “Applying Summarization Techniques for Term Selection in Relevance Feedback”, in Proc of the 24th annual international ACM SIGIR conf SIGIR '01, pp.01-09, Jan 2001.

[9] A. Singhal, “Modern Information Retrieval: A Brief Overview”, Computer Society Technical Committee on Data Engineering, Vol. 24, No. 4, pp. 35-42, 2001.

[10] B. Liu, "Web Data Mining: Exploring Hyperlinks, Contents and Usage Data”, Springer-Verlag, Berlin Heidelberg, 2002.

[11] V. E. Verelas and P. Raftopoulou, “Semantic similarity methods in Word-Net and their application to IR on the web”, in Web Information and Data Management, pp. 10–16, 2005.

[12] D. FRANK, "Comparing Rank and Score Combination Methods for Data Fusion in Information Retrieval", in Springer Science & Business Media The Netherlands, vol. 08, pp. 449–480, 2005.

[13] S.M.Shafi, and A. Rafiq, "Precision and Recall of five search engine for retrieval of Scholarly Information in the field of Biotechnologyfor Data Fusion in Information Retrieval", in Webology, vol.02, article 12, Aug 2005.

[14] C. Fellbaum,"Word Net(s)" Encyclopedia of Language & Linguistics, vol. 13, pp. 665–670, 2006,

[15] M. Song, Y. Song, X. Hu and B. Robert, " Integration of association rules and ontologies for semantic query expansion", Data & Knowledge Engineering, Vol. 63, No.1, pp 63-75 ,October 2007.