A Comprehensive Survey on Query Expansion Techniques, their Issues and Challenges

Neha Kathuria
Btech, CSE(Student)
Bhagwan Parshuram Institute of Technology, Delhi

Kanika Mittal
Asst Professor(CSE Dept)
Bhagwan Parshuram Institute of Technology, Delhi

Anusha Chhabra
Asst Professor(IT, Dept)
Bhagwan Parshuram Institute of Technology, Delhi

ABSTRACT

In order to improve the retrieval performance, the process of Query expansion is performed on the original user's query in order to reformulate the user's query. The basis of these query expansion techniques is to expand the query by adding the terms, which are in close proximity to the original query terms. Various query expansion techniques do not consider the context of the terms present in the user's query which can result in low precision and recall due to the ambiguity and vagueness of terms present in the query. Through this paper, a comprehensive survey is presented to study the various query expansion techniques proposed in literature by researchers and the various keyholes in the current scenario.

Keywords

Query Expansion, Natural Language Processing, Information Retrieval, Fuzzy Logic

1. INTRODUCTION

In today's times, Information Retrieval is considered as the major era of research. Its importance lies in the fact that it represents collection of documents, understanding and processing the query by the user and retrieval of the relevant documents. Information Retrieval is mainly used in retrieval of text documents from web. In Information retrieval, the stages are as follows: [1] user's information need, query formulation, user's query, document collection, filtering, indexing, document indexed, matching algorithm is applied, documents matched and document retrieved. The process of information retrieval is shown in fig 1. Various techniques are used for assessing the quality of IR [2]

Precision: - Percentage of retrieved documents pertaining to the query
Recall: - Percentage of pertaining documents and that are retrieved too while searching.

But precision and recall suffer from the disadvantage of being calculated for unordered set of documents. So, the concept of ranked retrieval came into existence that is the top k documents are retrieved. But still the problem associated is that people have different presentations while interacting with web [3]. Moreover, the problem that arises while dealing with information retrieval is word mismatch, spelling errors and usage of ambiguous words while typing queries. So, the solution to this problem is to expand the query as the incomplete and vague query makes it difficult for the search engine to retrieve the relevant documents. Query expansion is a technique used to expand the user's query which can be done either by thesaurus or automatically. Some of the Query expansion techniques are as follows:-Thesaurus based expansion, Wordnet based expansion, Query Logs based expansion, Implicit and Explicit Relevance Feedback and Pseudo Relevance Feedback. WordNet is the huge thesaurus used to get the lexical and semantic meaning of the word. But the drawback of using WordNet is that it is not possible to find the relationship between different parts of speech as in WordNet words are grouped on the basis of part of speech. To overcome this, WSD [5] is used in which the senses of words are determined using various algorithms to find the context the word is being used in query. Since user's queries are generally in the natural language and it contains imprecise and ambiguous words .To overcome this ambiguity and vagueness the fuzzy logic is designed by Zadeh in 1965 [6]. Fuzzy logic deals with imprecise and uncertain data so it acts as a good tool for information retrieval systems [7]. Fuzzy Logic is easy, flexible and more intuitive approach to deal with imprecise data as based on natural language. In fuzzy crisp sets are converted into fuzzy sets .Through the use of fuzzy, degree of membership is calculated. Fuzzy set model is used to define fuzzy queries. The fuzzy set contains linguistic variables. Each term of query is converted into fuzzy sets and degree of membership for each document is calculated. This calculation is done with the help of if-then Rule [8].

In this paper, we have discussed various approaches to expand the original query. In Section-II-V work done in the field of Query Expansion using different techniques is explained. In Section-VI, we have discussed some issues and challenges faced by different techniques used for Query Expansion. The article is concluded in Section VII.

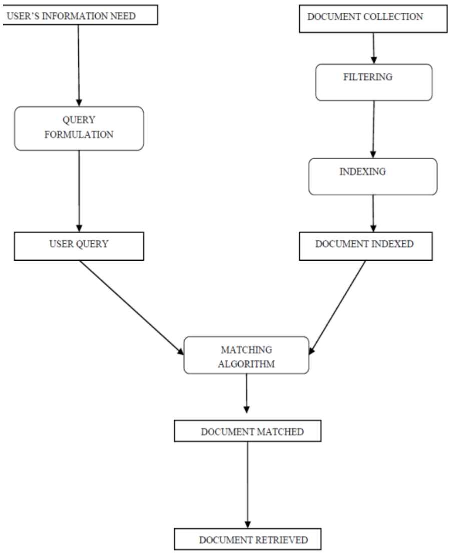

Figure 1: Process of Information Retrieval

Survey of Literature

Query Expansion [9] is a very useful technique to improve the query statements and retrieve relevant documents. The main aim of query expansion is to expand the query by adding an extra terms that are highly correlated to the query. This can improve precision and recall.

2. CONCEPT NETWORK BASED QUERY EXPANSION

In [10], Orland et al. proposed a method based on conceptual semantic theories in which the queries were expanded from concept network knowledge base. The query terms that matched in concept network, from there concepts are derived and new query terms are selected. And new query terms are selected. A directed graph model is constructed named Conceptual Word Cluster Space Graph (CWCSG) used to express semantic similarity among the concepts. Based upon CWCSG, user's query is extended to meet user's search needs accurately. This approach provides better performance as compared to synonyms dictionary of WordNet [11] but quality of this depends highly on the quality of concept network. In [12], relevant terms are weighed from top n documents and terms with higher weight are selected based on a threshold value. Concept is derived from those and phrases are identified. Then, the query terms and obtained phrases are matched to find additional query terms. G. Akrivas et al. have done Q.E. by adding semantic entities rather than terms. The semantic relations are explored in semantic encyclopedia to construct inclusion relation .The context of query is considered which is a fuzzy set of semantic entities. Also, this approach is integrated with user's profile [13]. In [14], the classification information is generated from the top ranked documents and the documents having the same classification information are grouped together to form clusters and the user can select the cluster corresponding to his needs.Chang et al. devised [15] a method of query expansion that is keyword based querying by combining query expansion and relevance feedback to get only concept-based information search.

3. QUERY EXPANSION BASED ON TERM WEIGHING

In [16] Lee et al. proposed a method in which documents are ranked by using domain ontology and user's profile. User's interests are captured and the query is expanded based on that. Then the expanded query is matched with retrieved documents to sort them according to the ranking. The performance of retrieval is enhanced in terms of MAP criterion (7%) by using pseudo-relevance feedback in [17] by Rahgozar et al. It assumes that the initial retrieval set of documents are relevant. Then, these documents are used to extract more relevant terms for query and re-weighting them. Since the top documents contain a closer user's context so similarity of top documents and weighing the set based on their context helped in efficient retrieval. Approximate matching of numerical terms too along with non-numerical terms is presented in [18] by Mittal et al.in which Fuzzy weighing of query terms is done with the help of fuzzy triangular membership function. Vector Space Model is used by Lin et al. [19] to represent both documents and queries. The fuzzy rules are framed to determine the degree of importance of relevant terms that is weights to get additional query terms. In this the precision and recall rates are increased to a greater extent. A.Hust et al. used the concept of similarity between new and old queries in [20]. From the documents that are relevant to old queries, the terms are extracted from them. So this approach uses global feedback approach for query expansion.

4. QUERY EXPANSION BASED ON WORD SENSE DISAMBIGUATION (WSD)

A query is expanded using WordNet and semantic relatedness in [21] by Li et al. measure modules. Also, Word sense disambiguation technique is used on query given by the user to analyze the sense of each term and context of query term. Now, the expanded query terms are generated from WordNet based on recovered concepts. This approach yields 7% precision improvement. In [22] Mittal et al. expands the query by finding similarity of ambiguous terms with other terms in query and assigning the weights to the similarity. An OWA (Ordered weighted averaging operator) is calculated for each sense of the target word and sense with highest similarity score is considered as the appropriate sense for that term. Combining this with the implicit feedback from the user, query is expanded .This results in better precision and hence optimization of query. Expansion of polysemy words by selecting those terms which are close to query terms with context meaning of terms is automatically incorporated in [23] by Tayal et al. in which the graph structure is created for the query terms. The relevant nodes which represent word senses are chosen from graph to be added to the query as additional terms to improve the retrieval performance. In [24] by Lapata et al., a graph based algorithm for large WSD is designed which identify most important node among the set of all nodes of graph. In their work, they have shown that how chosen lexicon and connectivity influences WSD performance. To get the context of particular concept and analyze the semantic relationships between them, WSD is applied along with thesaurus WordNet and Ontology of any domain in [25] by Valli et al. to retrieve the information. This method of query expansion improves the precision and recall and helps in retrieval of relevant information. In [26] Parapar et.al generated the queries as logical formulas using different connectives and types of linguistic information from WordNet.

5. QUERY EXPANSION BASED ON FUZZY LOGIC

A new method based on fuzzy logic is devised in [27] by Singh et al. in which the top-retrieved documents are considered as relevance feedback documents for finding additional terms. The weights of additional query terms are determined using fuzzy rules. Then, weights of additional and original query terms are used to form a new query vector and this vector is used to retrieve documents. This method of query expansion results in high precision rates, recall rates. In [28] Yogesh et.al had proposed a new fuzzy logic based ranking system which comprised of calculating term frequency, inverse document frequency and normalization. A Composite fuzzy logic is used to compute the relevance score of document with the query. Their method improves the overall performance of the retrieval in terms of precision and recall.

In [29] Ropero et al. have taken into account the other aspects of index terms that play an important role for determining term weights besides tf-idf; degree of identification of an object if the considered index term is used. The higher value of term weight implies more an index term identifies an object. This method is more efficient.

Martin and et.al refines the query with use of association rules. A fuzzy logic has been applied to check for appearance of term in retrieved documents (transactions) that has value between 0 and 1. Thus, the set of fuzzy association rules offers the choice of additional terms to be added to query to enhance the search efficiency in [30]. Huang et al. in [31] proposed a query expansion method using fuzzy association thesaurus in which the relevant terms are suggested to users and query is expanded by adding those terms and applying aggregation functions.

6. ISSUES AND CHALLENGES OF QUERY EXPANSION TECHNIQUES

From the above discussion, we have seen that the major tasks involved in query expansion expand the original query by adding relevant terms to the query by identifying the context meaning of terms present in the query. Despite of improving the retrieval process, query expansion techniques have some issues that needs attention and must be taken into consideration significantly. Some of the addressed issues are given in table1.

TABLE 1: Issues of Query Expansion Techniques

S.No	Query Expansion Techniques	Issues
1	Concept Based Network	1. Quality of conceptual query expansion depends on the quality of concept network. 2. Generally, useful for short query terms only. 3. Query interpretation is exhaustive. 4. Construction of Conceptual Word Cluster Graph is based on synonyms, dictionary which has probability of ambiguity
2	Term weighting techniques	1. Long documents tend to be fetched. 2. High tf-idf does not indicate that term is related to the user's query. 3. It even cannot capture semantics. 4. It is not able to consider the context of query.
3	Word sense Disambiguation	1. Using WSD is time consuming to recover correct senses of word. 2. Different dictionaries and thesauruses provide different divisions of words into senses so difficulty in disambiguating the exact sense.
4	Fuzzy Logic Based Approach	1. Associated weights of Fuzzy rules are not taken care of. 2. If fuzzy association thesaurus is used then no WSD is performed. 3. Calculation of Relevance Score is a tedious job using fuzzy logic.

7. REFERENCES

[1] A. Roshdi and A. Roohparvar, "Review: Information Retrieval Techniques and Applications", International Journal of Computer Networks and Communications Security Vol. 3, No. 9, pp. 373–377, 2015.

[2] R. Sagayam, S.Srinivasan and S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", International Journal of Computational Engineering Research (ijceronline.com), Vol.2, 2012.

[3] R.Kosala and H. Blockeel, "Web Mining Research: A Survey", ACM Sigkdd Explorations Newsletter-dl.acm.org, 2000.

[4] A. Kankaria, "Query Expansion techniques", Indian Institute of Technology Bombay, Mumbai, CSI Journal of Computing, Vol. 1, No. 2, 2012.

[5] R. Mihalcea and D. Moldovan, "Semantic Indexing using WordNet Senses", Proceedings of the ACL-workshop on recent advances in natural language processing and information retrieval: 38th Annual Meeting of the Association for Computational Linguistics-dl.acm.org, Vol.11, pp. 35-45, 2000.

[6] L.A.Zadeh, "Fuzzy Sets", Information and Control, Vol.8.3, pp. 338-353, 1965.

[7] A.Jain, K.Mittal, S. Sabharwal, "Information Retrieval in Fuzzy Logic Framework: A Survey", ICNICT, 2012.

[8] N.O.Rubens, "The Application of Fuzzy Logic to the Construction of the ranking Function of Information Retrieval Systems", Computer Modelling and New Technologies, Vol.10, No.1, pp. 20-27, 2006.

[9] Abdelmgeid Amin Aly, "Using a Query Expansion Technique to improve document retrieval", International Journal Information Technologies and Knowledge, Vol.2, 343, 2008.

[10] O.Hoeber, XD Yang and Y Yao, "Conceptual Query Expansion", International Atlantic Web Intelligence Conference, pp. 190-196, 2005.

[11] M.Peng, Q.Lin, Ye Tian, M.Yang, Y.Xiao and B.Ni , "Query expansion based on Conceptual Word Cluster Space Graph", Information Science and Service Science(NISS), 5th International Conference on New Trends, Vol.1, pp. 128-133, 2011.

[12] A.Jain, K.Mittal, S. Sabharwal, "Conceptual weighing Query Expansion on user profiles", National Conference on Communication Technologies & its impact on Next Generation Computing CTNGC Proceedings published by International Journal of Computer Applications (IJCA), 2012.

[13] G.Akrivas, M. Wallace, G. Andreou, G.Stamou and S. Kollias, "Context-Sensitive Semantic Query Expansion", Artificial Intelligence Systems (ICAIS) IEEE International Conference, pp. 109-114, 2002.

[14] Kang, J.W., H.K., Ko, M.C., Jeon, H.S., and Nam, "A Term Cluster Query Expansion Model Based on Classification Information in Natural Language Information Retrieval", Aritifical Intelligence and Computational Intelligence(AICI), IEEE, International Conference, Vol.2, pp. 172-176, 2010.

[15] C.H.Chang and CC Hsu, "Integrating query expansion and conceptual relevance feedback for personalized Web information retrieval", Computer Networks and ISDN Systems, pp. 621-623, 1998.

[16] GJ Hahm, MY Yi, JH Lee and HW Suh, "A Personalized Query Expansion Approach for Engineering document retrieval", Advanced Engineering Informatics 28 (4), pp. 344–359, 2014.

[17] P.Karisani, M.Rahgozar and F.Oroumchian," A Query term re-weighting approach using document similarity" Information Processing and Management,Vol.52, pp. 478-489, 2016.

[18] D. K. Tayal, S. Sabharwal, A.Jain and K.Mittal, "Intelligent Query Expansion for the Queries including Numerical Terms", Proceedings of International Journal of Computer Applications (IJCA), pp. 35-39, 2012.

[19] HC Lin, LH Wang and SM Chen, "Query Expansion for Document Retrieval by Mining Additional Query Terms", Information and Management Sciences, 19(1), pp. 17-30, 2008.

[20] A.Hust, S.Klink, M. Junker and A. Dengel, "Query Expansion for Web Information Retrieval", GI Jahrestagung, pp. 176-182, 2002.

[21] J Zhang, B Deng and X Li, "Concept Based Query Expansion Using WordNet", Proceedings of the 2009 International e-Conference on Advanced Science and Technology IEEE Computer Society Washington, DC, USA, pp. 52-55, 2009.

[22] K. Mittal and A. Jain, "A graph Based Query Expansion using Semantic Similarity and OWA Operator" , ICTACT Journal on Soft Computing, Vol.5 pp. 896-904, 2015.

[23] A. Jain, K. Mittal and D. K. Tayal, "Automatically incorporating context meaning for query expansion using graph connectivity measures", Progress in Artificial Intelligence, Vol.2, pp. 129–139, 2014.

[24] R.Navigli and M. Lapata, "An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation", IEEE transactions on pattern analysis and machine intelligence, VOL. 32, NO.4, pp. 678-692, 2010.

[25] M.Barathi and S.Valli, "Ontology Based Query Expansion Using Word Sense Disambiguation", (IJCSIS) International Journal of Computer Science and Information Security, Vol.7, No. 2, 2010.

[26] D.Parapar, Á. Barreiro and D.E. Losada, "Query Expansion Using WordNet with a Logical Model of Information Retrieval", IADIS AC, pp.487-494, 2005.

[27] J. Singh and A.Sharan , " A new Fuzzy Logic based Q.E. Model for efficient IR using Relevance Feedback Approach", Neural Computing and Applications, pp. 1-24, 2016.

[28] Y.Gupta, A.Saini and A.K. Saxena, "A new Fuzzy Logic Based Ranking Function for efficient IR System" Expert Systems with Applications, 42(3), Vol.42, pp. 1223–1234, 2015.

[29] J. Ropero, A. Gomez, A. Carrasco, C. Leon and J. Luque , "Term Weighting for Information Retrieval Using Fuzzy Logic", 2012.

[30] M.J. Martin-Bautista, D. Sanchez, J. C. Martinez, J.M. Serrano and M.A Vila, "Mining web documents to find additional query terms using fuzzy association rules", Fuzzy Sets and Systems,148(1), pp.85-104, 2004.

[31] H.M Lee, S.K Lin and CW. Huang, "Interactive Query Expansion Based on Fuzzy Association Thesaurus for Web Information Retrieval", Fuzzy Systems, 10th IEEE International Conference on Vol.2, pp. 724-727, 2001.