A study: query expansion methods in information retrieval

Journal of Physics: Conference Series **1367** (2019) 012001 doi:10.1088/1742-6596/1367/1/012001

A study: query expansion methods in information retrieval

Lasmedi Afuan[1], **Ahmad Ashari**[2], **Yohanes Suyanto**[3]
[1]Department of Informatics, Universitas Jenderal Soedirman, Indonesia
[2,3]Department of Computer Science and Electronics Faculty of Mathematics and Natural Sciences Universitas Gadjah Mada, Indonesia

E-mail: lasmedi.afuan@unsoed.ac.id[1], ashari@ugm.ac.id[2], yanto@ugm.ac.id[3]

Abstract—This paper study about a review of the literature on Query Expansion (QE) methods. The main aspects of the review study are the methods, limitations of the QE methods, and datasets used in QE. As the results of this study, we can conclude several methods used on QE i.e., Ontology, Association Rules, Wordnet, Methathesaurus, Synonym mapping, Concept-based, Local Co-occurrence, and Latent Semantic Indexing (LSI). Each of the methods still has limitations. For datasets used in QE, many researchers use public datasets.

1. Introduction

The abundant number of documents on the Internet creates a problem for the user. The user has difficulty finding relevant documents or information to their needs. Information Retrieval (IR) is required to retrieve related documents to user's queries. IR is the process of finding data in the form of text by the required information from a collection of documents stored on the computer [1]. IR provides information on the subject matter required. The data includes text, audio, video, and other documents. IR aims to produce documents relevant to the user's queries in a short and precise time.

The current IR research comes with two significant developments: how to index documents and how to retrieve relevant documents to user queries [2]. IR research conducted at different levels but with the same objective to improving the relevance of the document, such as [3] adapting the classical VSM model to ontology-based IR, [4] proposed ontology-based IR, [5] proposed IR using PSO and IR with QE [6], [7], [8], [9], [10]. IR research has been proposed, generally using keywords in searching the document content [4]. Keywords used by users are often the same word but have different meanings, and different words have the same meaning. Also, keywords entered by users are too short cause ambiguity [11].

The highest percentage of word lengths used by users in a query is one to three words. Meanwhile, defining short queries is a query that is less than four words [12]. Users are also unable to represent the required information needs into the queries. As a result, documents generated by IR do not match the user's needed. The number of relevant documents generated depends on the query of the user's queries. The vocabulary of user queries that mismatch and miss concept with document also causes no document to be retrieved [12].

The good IR should be able to bridge potential distances between documents and user queries [12], To overcome these problems, research in IR proposes many solutions, one of which is with QE [13]. QE is believed to be able to overcome the problems associated with user query representation. The approach is used to overcome the problem in the ineffectiveness of document retrieval by modifying the query to improve the quality of the user query, which believed that the less accurate query is the main problem related to the relevance of the document to IR [14].

This study reviews the literature on QE. The main contributions of this paper and differences with other review papers are in this review study conducted on aspects of methods, limitations, and datasets used in QE. Meanwhile, other review papers only review the methods and tools for QE.

The remainder of this paper is organized as follows. In section 2, challenges in information retrieval systems (IRS). In section 3, We are going to discuss QE method used for improving IRS. A small summary and further study of this area will be concluded in section 4.

2. Materials And Methods

The review questions (RQ) are specified to keep the review focused. The review questions addressed by this literature review are:

- (RQ1): What kind of methods are used for QE?
- (RQ2): What are the limitations of methods used in QE?
- (RQ3): What kind of datasets is the most used for QE?

The reason for the review question above is to find out the methods, the limitations of the methods used in query expansion, and the dataset commonly used in QE. Because in general, several reviews are only limited to the methods used in QE. The search process for literature consists of several activities, such as selecting the digital library, defining the search string, executing the search string, and retrieving a list of primary studies from digital libraries. Digital libraries used in the search process, i.e., ACM digital library, IEEE, and Elsevier.

3. Challenges in Current Information Retrieval Systems (IRS)

A lot of research on Information Retrieval (IR) has been proposed, based on the literature there are several models of classical IR, i.e., Boolean Model, Vector Space Model, and Probabilistic Model [15]. The current IR research comes with two significant developments: how to index documents and how to retrieve documents [2]. Currently, IR still has a complex problem, which relates to two subjectivities, namely the relevance of user needs and the uncertainty and vagueness that characterize the document retrieval process [15].

The relevance of documents retrieved from IR is still a major issue. The ineffectiveness of information retrieval systems often caused by query inaccuracy. Retrieve information from the Internet using information retrieval systems often need precise keywords from multiple fields to achieve the best result. Information retrieval systems often need the exact keywords to return a high-quality result list. Hundreds of thousands of irrelevant documents will be returned if the selected keywords are too general. This has become a problem for a user when they are not sure about the nature of the content they needed or the difficulties of describing the nature of the context of the necessary information in just a few keywords.

The retrieved irrelevant document is also caused by user queries that mismatch and miss concept with document collection. Often, the same word has a different meaning, and different words have the same meaning. IR must be able to bridge the user query that mismatch and miss concept. Several approaches have been proposed to address problems with user queries, such as query rewriting, query suggestion, and QE. Research on the use of QE shows an increase in IR performance.

4. Discussion

4.1. Query Expansion Methods

Several researchers have researched QE. The latest research [10] focusing on QE with an ontology approach that combines the intensive expansion, extensional expansion, and word refinement approaches. Research conducted by [16] proposed the use of the Latent Semantic Indexing (LSI) method. This method is potent, implemented on two types of algorithms, namely the Singular Value Decomposition (SVD) and Probabilistic LSI. LSI builds semantic spaces, mapping each term into

space, and grouping automatically based on the meaning of the word. But, with the LSI method, it is difficult to control the degree of QE, and it could be the expansion query contains many irrelevant terms. To overcome this, [17] proposed expansion with Local Co-occurrence approach, this approach based on the frequency of word occurrences in document collections. This method can increase the effectiveness of IR in the range of 6 to 13%. But, this method was unable to display connectedness and meaning of the word.

The research conducted by [18] proposed QE through term selection in the relevance feedback process based on the Rocchio formula on the return of XML document information. This approach can overcome two major problems in the retrieval of XML document information that is the problem of overlapping the elements are taken and the problem of irrelevant elements. It's just that the use of relevance feedback depends heavily on user judgment, whether the resulting document is relevant or not. Thus, if the document is considered relevant but it is not, the IR result is less relevant. Similar to the LSI approach, relevance feedback has not been able to show the connectedness and meaning of words.

Research conducted by [19] proposed QE using ULMS Metathesaurus. A word or user query mapped into UMLS CUIs using Meta Map, then MRCONSO Metathesaurus table identifies synonyms of words, and those words used for expansion queries. But, the use of Metathesaurus in some user queries, the expansion query degrades the performance of IR. The study conducted by [8] also uses WordNet to search for synonyms of words entered by the user. The query expansion process is done by identifying Part of Speech (POS) of every word using Tagger POS. Then after that, synonymous identification of each word to expand the query using WordNet. The results of this study indicate an increase in precision and recall of about 40% and 24% compared with no expansion query.

Research conducted [20] is similar to the research conducted [19] performs the query expansion by mapping the word and searching for synonyms of the word entered by the user, relevant searches are retrieved and rewrote. The research conducted by [21] proposes two stages in the QE method that is to reduce the weighting of the word that is too much (out weighting) by classifying the term on the query based on semantic relationships, then using the recursive structure of Hopfield's most related network in other words selected. For word candidate extraction using WordNet. The evaluation results using CACM and CERC collections showed an increase of 4% - 12% using MAP. But, the use of WordNet/Metathesaurus in some user queries, the expanded query degrades the performance of the IR, but it is also less able to display interrelationships between words. To overcome the interconnectedness of words in QE, [22] proposed a QE using Association Rules between terms, using the 95 SDA collection dataset. [23] proposed expansion queries on retrieval patents based on domain lexicon. But, this approach is unable to display a relationship between concepts.

The study conducted by [24] proposes a QE based on the concept of using semantic connectivity through indirect graph concepts. Research conducted by [9] proposed QE with Thesaurus MeSH (The Medical Subject Headings) ontology to improve IR on medical data collection. By combining an independent subsystem to retrieve textual and visual information. The study used the dataset provided by CLEF that is imageCLEFmed 2005 and imageCLEFmed 2006. The IR effectiveness test measured by MAP. The results of the study concluded that the use of MESH ontology and QE could increase IR not only textual but also visual.

The research [10] proposed QE with ontology using a hybrid approach by combining intensive expansion, extensional expansion, and word refinement. In the first step, the intentional module takes the term of ontology meaningful to the word entered by the user. Then, the word added to the query candidate list. Next, the search module is executed, generating a document related to the query list of the intensive module. [25] proposed ontology to search for the semantic similarity of words. The user keyword is extracted and then searches on ontology ten lists of words that have semantic similarities. Next, Boolean operators "AND" and "OR" are used to create the same semantically new query with the initial query. But, the use of ontologies is unable to display the interconnection between words in the document. Based on a review of literature review, QE research grouped into seven methods that are often used to perform QE as presented in Table 1 such as LSI method, Local Co-Occurrence, Relevance Feedback, Concept-Based, WordNet, Synonym Mapping, Association Rules, and Ontology.

4.2. The limitations of QE methods

Based on our review study, the methods used in QE have limitations. In the study using ontology in QE, the method is unable to display the relationship between the terms in the document. Meanwhile, the Association Rules are limited to capture the relationship between concepts. The use of synonym mapping depends heavily on the completeness of the dictionary used and can occur over expansion, causing no relevant documents for some queries. The limitations of QE methods shown in Table 1.

Tabel 1. QE Methods

No	Cites	Uses Methods	Limitations
1	[10]	Ontology	a
2	[25]	Ontology	a
3	[9]	Ontology	a
4	[26]	Ontology	a
5	[27]	Ontology	a
6	[28]	Ontology	a
7	[7]	Association rules	b
8	[22]	Association rules	b
9	[23]	Lexicon domain	c
10	[13]	Association rules	b
11	[20]	Synonym mapping	d
12	[8]	WordNet	e
13	[21]	WordNet	e
14	[19]	Metathesaurus	f
15	[24]	Concept-Based	g
16	[18]	Relevance feedback	h
17	[17]	Local co-occurrence	i
18	[16]	Latent Semantic Indexing (LSI)	j

Based on Table 1 columns limitations, we describe in this paragraph a) Cannot display the relationship of the appearance of words in a document; b) Cannot capture the relationship between terms concept in a document; c) Can not obtain the relationship between terms concept in a document; d) It depends on the dictionary and can make over expansion in some queries; e) The use of WordNet on some user query, it can reduce the performance of IR; f) The use of Metathesaurus on some user query, it can reduce performance of IR; g) cannot display the connection between the appearance of words and meanings in the document; h) i) Unable to understand the meaning of connectedness of words; j) It is difficult to control the degree of expansion query and the expansion query contains irrelevant terms;

4.3. QE Dataset

Investigation results of our studies towards datasets, it has been used to perform QE shown in Table 2. The previous research used two types of datasets to conduct QE, i.e., Private and Public datasets. Based on our observation towards literature, QE uses various datasets, i.e., TREC, Wikipedia, DBpedia, CLEF, etc. Based on Table 2, we can conclude that fourteen of eighteen research studies use public datasets in their studies and other studies use private datasets.

Tabel 2. QE DATASET

No	Cites	Dataset	Available
1	[10]	TREC-CDS	Public
2	[25]	TREC and CLEF	Public
3	[9]	ImageCLEFmed	Public
4	[26]	Wikipedia	Public
5	[27]	TREC	Public
6	[28]	TREC	Public
7	[7]	DBPedia, Wikipedia	Public
8	[22]	CLEF 2003 corpus	Public
9	[23]	CLEF-IP patent	Public
10	[13]	CISI	Public
11	[20]	Malayalam	Private
12	[8]	Code	Private
13	[21]	CACM,CERC	Public
14	[19]	Medline	Public
15	[24]	Wikipedia	Public
16	[18]	XML	Private
17	[17]	TREC	Public
18	[16]	CA	Private

5. Conclusion And Future Work

In this paper, we have reviewed the methods, method limitations, and dataset used by QE on IR. As the results, there are seven methods that can be used in QE such as Ontology, Association Rules, Wordnet, Methathesaurus, Synonym mapping, concept-based, Local Co-occurrence, and Latent Semantic Indexing (LSI). From the literature review, QE Research uses many public datasets. The datasets used in QE include TREC-CDS (Clinical Decision Support), TREC, CLEF, I'm-ageCLEFmed, Wikipedia, DBpedia, Wikipedia, CLEF 2003 corpus, CLEF-IP patent, CISI Dataset, Malayalam, Code, CACM, CERC, Medline, XML, and CA. The methods used in QE still have limitations. For future work, to overcome these limitations, this study proposes a combination of several methods, such as the use of Ontology and Association Rules, to overcome the limitations of each method.

References

[1] C. D. Manning, P. Raghavan, and H. Schutze, 2009, "An Introduction to Information Retrieval," *Online*, no. c, p. 569.

[2] B. M. Sanderson and W. B. Croft, 2012, "The History of Information Retrieval Research," in *IEEE*, vol. 100, pp. 1444–1451.

[3] P. Castells, M. Fernandez, D. Vallet, M. Fernandez, and D. Vallet, 2007, "An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 2, pp. 261–272.

[4] M. Fernandez, I. Cantador, V. Lopez, D. Vallet, P. Castells, and E. Motta, 2011, "Semantically enhanced Information Retrieval: An ontology-based approach," *Web Semantics: Science, Services, and Agents on the World Wide Web*, vol. 9, no. 4, pp. 434–452.

[5] A. Gomathi, J. Jayapriya, G. Nishanthi, K. S. Pranav, and P. K. G, "Ontology Based Semantic Information Retrieval Using Particle Swarm Optimization, 2015," *International Journal on Applications in Information and Communication Engineering*, vol. 1, no. 4, pp. 5–8.

[6] D. Zhou, S. Lawless, J. Liu, S. Zhang, and Y. Xu, 2015, "Query Expansion for Personalized Cross-Language Information Retrieval," *International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*.

[7] M. Amina, L. Chiraz, and Y. Slimani, 2016, "Short Query Expansion for Microblog Retrieval," *Procedia - Procedia Computer Science*, vol. 96, pp. 225–234. [Online]. Available: http://dx.doi.org/10.1016/j.procs.2016.08.135

[8] M. Lu, X. Sun, S. Wang, D. Lo, and Y. Duan, 2015, "Query Expansion via Wordnet for Effective Code Search," *IEEE*, pp. 545–549.

[9] M. D. Galiano, M. M. Valvidia, and L. U. Lopez, 2009, "Query expansion with a medical ontology to improve a multimodal information retrieval system," *Computers in Biology and Medicine*, vol. 39, pp. 396–403.

[10] J. Choi, Y. Park, and M. Yi, 2016, "A Hybrid Method for Retrieving Medical Documents with Query Expansion," in *Big Data and Smart Computing (BigComp)*, pp. 411–414.

[11] L. Araujo, H. Zaragoza, 2010, J. R. Pérez-agüera, and J. Pérez-iglesias, "Structure of morphologically expanded queries: A genetic algorithm approach," *Data & Knowledge Engineering*.

[12] D. Pal, M. Mitra, and S. Bhattacharya, 2015, "Exploring Query Categorisation for Query Expansion: A Study," *CoRR*, pp. 1–34.

[13] A. Abbache, F. Meziane, G. Belalem, and F. Z. Belkredim, 2016, "Arabic Query Expansion Using WordNet and Association Rules," *International Journal of Intelligent Information Technologies*, vol. 12, no. 3.

[14] J. Ooi and H. Qin, 2015, "A Survey of Query Expansion, Query Suggestion, and Query Refinement Techniques," *International Conference on Software Engineering and Computer Systems (ICSECS)*, pp. 112–117.

[15] S. Marrara, G. Pasi, and M. Viviani, 2017, "Aggregation operators in Information Retrieval," *Fuzzy Sets and Systems*, vol. 1, pp. 1–17. [Online]. Available: http://dx.doi.org/10.1016/j.fss.2016.12.018

[16] S. Deerwester, S. T. Dumais, and R. Harshman, 1990, "Indexing by latent semantic analysis," *Journal of the American Society for information science*, vol. 41, no. 6, pp. 391–407.

[17] M. Mitra, C. Buckley, and F. Park, 1998, "Improving Automatic Query Expansion," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*.

[18] M. Mataoui, F. Sebbak, F. Benhammadi, and K. B. Bey, 2015, "Query Expansion in XML Information Retrieval A new Approach for terms selection M'hamed," in *Modeling, Simulation, and Applied Optimization (ICMSAO)*, pp. 4–7.

[19] M. R. A. Nawab, M. Stevenson, and P. Clough, 2015, "An IR-based Approach Utilising Query Expansion for Plagiarism Detection in MEDLINE," *Journal of Computational Biology and Bioinformatics*, vol. 5963, no. APRIL 2015, pp. 1–9.

[20] A. Babu and S. L, 2015, "An Information Retrieval System for Malayalam Using Query Expansion Technique," *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1559–1564.

[21] A. Noroozi and R. Malekzadeh, 2015, "Integration of Recursive Structure of Hopfield and Ontologies for Query Expansion," *International Symposium on Artificial Intelligence and Signal Processing*.

[22] A. Bouziri, C. Latiri, E. Gaussier, and Y. Belhareth, 2012, "Learning Query Expansion from Association Rules Between Terms,".

[23] F. Wang and L. Lin, 2016, "Domain Lexicon-based Query Expansion for Patent Retrieval," *International Conference on Natural Computation*, pp. 1543–1547.

[24] A. Boubacar and Z. Niu, 2013, "Concept Based Query Expansion," *International Conference on Semantics, Knowledge, and Grids*.

[25] H. Al-chalabi, S. Ray, and K. Shaalan, 2015, "Semantic based Query Expansion for Arabic Question Answering Systems," *First International Conference on Arabic Computational Linguistics Semantic*, pp. 131– 136.
[26] M. Farhoodi, M. Mahmoudi, A. Mohammad, Z. Bidoki, A. Yari, and M. Azadnia, 2009, "Query Expansion Using Persian Ontology Derived from Wikipedia," *World Applied Sciences Journal*, vol. 7, no. 4, pp. 410–417.
[27] Q. Jin, J. Zhao, and B. Xu, 2003, "Query expansion based on term similarity tree model," *International Conference on Natural Language Processing and Knowledge Engineering*, pp. 400–406.
[28] R. Mandala, T. Tokunaga, and H. Tanaka, 1999, "Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion," *Proceedings of the 22Nd Annual International Acm Sigir Conference on Research and Development in Information Retrieval*, pp. 191–197.