IJCSN - International Journal of Computer Science and Network, Volume 7, Issue 4, August 2018
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

A Review of Query Expansion Approaches

[1] Hiba AL.Marwi; [2] Mousa Ghurab

[1] Computer science, Sana'a University
Sana'a, Al-Yemen

[2] Computer science, Sana'a University
Sana'a, Al-Yemen

Abstract - Since the content of the web is increasing rapidly, retrieve most relevant document to user need is difficult task. Further more user query in most cases is vague and formed by a few keywords. Sometimes, user query is well formed, but in different form as the document collection. A number of approaches have been used to process user query, and enhance its performance. One of the most successful approaches is query expansion, whereby additional terms are added to the original query. In this paper, we attempt to introduce better understanding of query expansion approach by reviewing and comparing the effectiveness of these approaches. The following questions are answered. What are the different types of query expansion approaches? What are the advantages and disadvantages of each type?

Keywords - Information Retrieval; Query expansion; Word embedding; WordNet.

1. Introduction

Using the internet is becoming one of the most important parts of our daily life. One of the common uses of the internet is to find information on a specific topic. As a result of a rapid expansion of available information presented to us on the internet and the huge lexical gap between user query and his intent, the process of retrieving documents which match the user query has become a difficult task [1, 2]. Although many researchers have been done to improve the efficiency of information retrieval, there are still many problems yet unsolved. As David Seuss states: "Ten Years into the web and the Search Problem is Nowhere Near Solved" [3]. A number of researchers [4-6] have reported that the average query length was only around 2.30 words which may not be enough to convey the concept that the user is looking for [7]. One of the most important problems facing information retrieval IR is Vocabulary mismatch which is known as vocabulary problem [7, 8]. This may be caused by the polysemy or synonym, the polysemy (different words with the same meaning) such as 'car' and' automobile' which causes decrease in precision, On the other hand, Synonym is a word having the similar or nearly the same meaning such as: apple which causes decrease in Recall.

According to [78] 32. 5% of users revised their query and 29. 3% added more than one keyword to research. Due to a ll these reasons, several approaches have been introduced i n the user query processing field (word sense disambiguati on, relevance feedback, query refinement, interactive quer y refinement, and search results clustering and re-ranking). Although these approaches are effective, sometimes they a re not able to provide accurate information to the user. One of the well-known and successful techniques is query expa nsion (QE) proposed by [9]. This technique not only expre sses the user's need more precisely by modifying original q uery but also removes the ambiguity exists in user's need a nd hence improves the performance [10]. This article is a c omprehensive study which reviews a major query expansio n approach from the oldest to latest one and discusses their performance.

2. Query Expansion Approaches

When the users submit their query **Q**, IR system modifies the query to improve its quality by computing the importance of each term that occurs in the query and document d to generate new expansion query **Q'**. The similarity sim ('q, d) between query **'q** and document **d** can be computed as Eq. (1)

$$\text{Sim}\,(\bar{q}, d) = \sum_{t \in \bar{q} \cap d} \overline{w}_{t,\bar{q}}.w_{t,d} \qquad (1)$$

Where $w_{t,\,q}$ and $w_{t,\,d}$ are the weights of term t in query q' and document d. according to weighting function.

The information retrieval approaches that have been used may be classified on the basis of conceptual paradigm used for finding the expansion terms as being either based on st

atistical, Linguistic analysis, semantic, and hybrid (based o n statistical and semantic) approaches. The detailed taxono my of query expansion approaches is shown in figure 1.

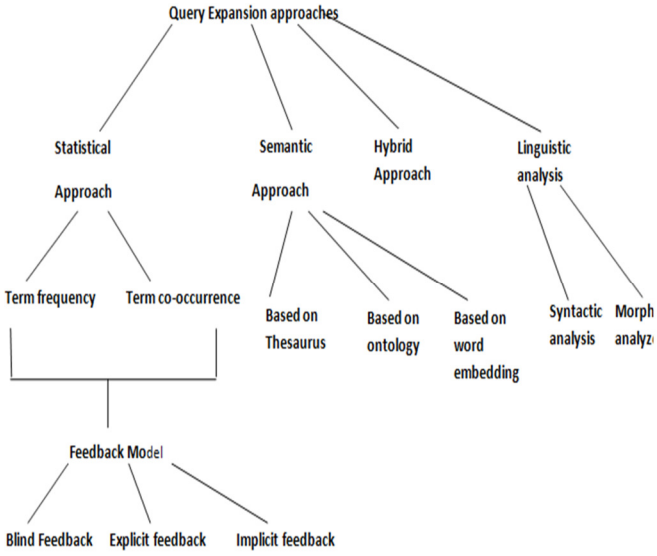

Fig. 1. Taxonomy of approaches.

3. Statistical Approach

Usually it depends on calculating the number of terms occ urring in the document and choosing the most frequent one s as most likely terms related to the query. This approach i s efficient but it does not work well in short text [57]. Also it negatively affects on the IR performance because it does not take care about semantically similar terms and it needs high computational cost. We can further classify statistical approach into:

3.1 Term Frequency (Unigram Modeling) (full I ndependence):

According to the Information Retrieval (IR) theory, both document and query represent using "bag of word" model where a document d and query q represent as vector space [11]. Each weight in the vector represents the importance of word to the document as whole, in the same way for query. The weight is computed in different ways. For instance, in the term frequency (TF) model, the weight is computed by considering the frequency of terms in a document. Researcher utilized two cut-off points. These cut-off points separated the ranked list of terms in three parts. First part upper cut-off contains very frequent terms; second part contains terms which were neither too rare nor too frequent; third part lower cut-off contains (very infrequent terms those do not contribute significantly in documents). Luhn [12] Considered only second part and named his proposed approach as resolving power technique. Fang and Zhai [13] Found that TF often leads to poor results due to possible negative weights for certain frequent terms. Therefore, the number of documents containing term (dft) and the number of occurrences of term t in dataset (cft) used to determine the weight. Jung and Park [14] Proposed weighting method that not only considers occurrence terms but also absent terms which are negatively weighted. Efron [15] Extended term frequency model by trace the behavior query terms while the collection of documents changes over time.

Although this model achieves good performance, it does not work well in all cases as it depends on exact matching of terms (does not consider semantic information). Furthermore, it suffers from data sparsity. Cao [16] showed that opted expansion terms depend on study of terms distribution is not good enough as most of those terms unrelated to the query and harmful to the retrieval. In order to perform semantic term matching El-Mahdaouy, El Alaoui, and Gaussier [17] have proposed to enhance term frequency based on distributed representation of word. The obtained results on Arabic TREC 2001/2002 showed that word Embeddings improve significantly the term frequency. Gao [83] proposed a novel query expansion approach for Community Question Answering based on term frequency.

3.2 Term Co-occurrence Approach:

Few years later, Attar, Fraenkel and Bai [18, 19] believed t hat if two terms appears in the same document that means i t is semantically related to each other. From this view, som e studies compute co-occurrence between each term in the documents and the terms of query. If there is a strong corre lation between them, the term will be selected as expansio n features. One downside of this approach co-occurrence b etween terms in document and query does not necessarily mean the expansion terms are correlated to meaning of the query as whole. Moreover, good expansion terms are not n ecessary co-occurrence with query word. Few years later, Bai [20] addressed this weakness by instead of study co-oc currence between query terms and documents. He suggeste d studying co-occurrence between terms in same documen t, but it does not take care about the position of terms (the t erms which appear in the same sentence seem more correla ted than the terms appear distantly with document). There are several approaches to find co-occurrence between term s. Shaalan, Al-Sheikh, and Oroumchian [21] have used Ex

pectation Maximization (EM) algorithm for selecting most related terms to user query. Bai [20] used association rule t o find co-occurrence between terms but it does not take car e about dependency between them. Wei and Bressan [22] i mproved the effectiveness of association rule by consideri ng co-occurrence frequency, confidence and direction of th e association rules. Momtazi and Khudanpur [53] propose d to find similarity between terms in all sentences in the co rpus. The corpus must be syntactically parsed if there is a s yntactic relation between them, the co-occurrence can be d efined. Probability approach is one of the most promising approaches to address this issue which builds a model base d on the probability distribution of term where the highest probabilities often selected as expansion terms [7]. All the above statistical method uses one of the following feedbac ks Model as a source for expansion:

4. Relevance Feedback (Local Approach)

One of the most well-known approaches is the Pseudo-Relevance Feedback (PRF). It selects set of useful terms from top k documents retrieved in response to the original query [23-27]. It may slower at runtime and require extra search and space, but it considers sufficient Techniques [2, 28, and 29]. It Moves the results to the most popular meaning of the query. There are three different types of feedback: ad hoc or blind feedback, implicit feedback and explicit feedback.

4.1 Blind Feedback:

One of the effective ways to reformulate user query is called Blind feedback [30]. It is based on the assumption that the top retrieved document is most relevant to user query, but in most cases those documents are irrelevant. Using those documents as a source for expansion generates expanded query more similar to the retrieved documents, rather than user intention [7] and may drift query to another topic because those documents may fill in different topics such sport or health [31]. As a result, it gains high recall with loss precision [32]. A further reason for decreased precision is the low dimensionality of user query beside high dimensionality of the feedback document. As a consequence, the performance for the expanded query is not better than the original one [33]. Several approaches have been proposed to improve the PRF performance. Sakai, Manabe, and M. Koyama [34] Proposed that instead of using all the top k documents, we could choose a subset of which is likely to be highly relevant. Cao [16] showed that using subset terms from pseudo relevance document can be used to avoid query drift problem. Lv, Zhai [24] Proposed a novel term weighting function which assigns high weight to the terms in the document which are closer to the query terms. Chinnakotla [79] proposed Multilingual PRF where the original query is written in L1 translated to other language L2. PRF is worked on both collections of L1 and L2. The feedback model for query in L2 (FB2) is translated back to L1. This feedback model (FB2) is combined with the feedback model (FB1) for query in L1. This approach gives a better performance as it considers both co-occurrences based terms as well as semantically and lexically related terms consideration. Attar and Fraenkel [18] built clusters from top document to use as a source for expansion. Although this study increases precision, it does not take into account the response time. However, IR system considers precision, recall and fast response time.

4.2 Explicit Feedback

Using blind feedback as a source for expansion is less effe ctive due to the most of irrelevant documents. Explicit fee dback overcomes this limitation by allowing users to provi de feedback determining the relevance of a documents usi ng Graded or binary feedback [35, 36]. In relevant feedbac k, the users mark the documents as relevant or irrelevant to original user query. Graded relevant feedback shows the re levance of document by using specific rate such as: numbe rs, letters.

4.3 Implicit Feedback:

Although explicit feedback approach is sufficient, it is tim e consuming and adds an additional burden for the user. It makes expansion process less applicable [37]. As a conseq uence, most IR systems use implicit feedback where the sy stem automatically determines the relevance of documents by observing user behavior whether considers duration of t ime that the user spent in viewing a document, selecting do cument or not, any scrolling action[38]. In fact, a full auto matic approach exhibits a low performance especially whe n a query is too ambiguous. As a result, Okabe and Yamad a [39] proposed hybrid method which uses minimal explici t feedback to identify implicitly other documents related to user query. Most researches in the IR flied built relevance model from a given ad-hoc collections such as TREC due t o its small size. Researcher in [40], Proposed exploit mixtu re relevance models [MORM] which is a simple relevance model focuses in both target and external collection. The major drawback of this relevance model is the difficulty of determining the contribution of external collection in feedb ack model. To overcome this limitation, Weerkamp, Balo

g, and de- Rijke [41] built EEM (external expansion mode l) from external collection. To improve the performance of EEM and generate more accurate documents, Liu and Jung [42, 43] proposed a method to effectively using the externa l collection based on clustering the external collection usin g k-means clustering algorithm. A lot of similar studies cre ate cluster using different approaches such as [44].

5. Linguistic Analyses:

In order to understand the exact intent of a user query, que ries may be analyzed based on their linguistic prosperities. However, previous research has indicated that various Lin guistic analysis in terms of morphological, syntactical anal yzes are a challenge task as it involved much depth and fac ed Complexity in Breaking down words.

5.1 Morphology Analyzes

The purpose of a Morphological analyzer is to examine the structure of the word, retrieve grammatical features of a m orphologically inflected word. Early studies on IR have fo cused on morphology analyze of document, hence many ef forts have been focused on developing a stemmer such as [45-47]. Those stemmers depended on set of rules and use d lookup table for finding the root. This technique can be c lassified into two main categories: root-based and stem-bas ed approach. The root-based approaches reduce derived w ords to their root by removing prefixes and suffixes, and th en extracting the root of the stripped form, using a list of p atterns. The stem-based approaches namely light stemmin g, truncates the stem from the words by removing affixes u sing a list of prefixes and suffixes. Root-based stemmers m ay group none semantically similar words into the same w ord index (root) while light stemming techniques may fail t o group two semantically similar terms to the same stem. Al-Serhan and Ayesh [48] have tackled this drawback by u tilizing neural network to extract root. It significantly incre ases the IR performance. Most stemming approaches intro duce a large amount of noise in documents. Elayeb and Bo unhas [49] have explained the limits of morphology analyz ed in Arabic IR.

5.2 Syntactic Analysis:

One of the most useful approaches is a syntactic analysis which aims to find expansion terms for user query terms fr om top ranked document by inducing the most relevant pat hs to user query from parse trees [50]. It is less efficient ap proaches because it requires exact match.

In natural language, the user query made up of terms interc onnects based on grammatical relation (syntactical structur es). If considered properly, user query can be reformulated into meaningful query terms which can be used as query e xpansion terms.

Queries consist of multiple terms each of which plays a dif ferent role. Some of them represent query content while ot hers connect query terms. Therefore, it is necessary to iden tify the role of each term in order to extract expansion con cepts. Traditional approaches determine the role of each te rm in query based only on part-of-speech which causes mi ss-labeling role assignment. Punyakanok and Roth [51] uti lized syntactic parsing to assignment role for each term de pending on the grammatical relations between query term s. In most cases the parser (linguistic parser) fails to identif y most precise relations between a pair of terms which gen erates relation tagged as unidentified (untagged terms prob lem). Selvaretnam and M. Belkhatir [52] handled this issue by utilizing term frequency to determine the type of role to untagged terms where the assumption assigns the most fre quently occurring terms as key terms. Syntactic processing is a successful approach, but sometimes it prunes errors du e to semantic ambiguities on user query.

6. Semantic Approaches

In the last years because of the best of understanding the limitations of statistical approach, Semantic approach has been presented to overcome these limitations [54]. The basic idea behind this approach is to find the expansion terms based on the representative meaning of the query in its context rather than matching single terms. This approach tries to remove ambiguity from query by discovering its meaning and understanding contextual meaning of terms in searchable data space [55]. We classify this method into:

6.1 Standard Thesaurus Based (Global approaches):

It is a well-known approach which aims to rephrase query based on its context by adding new words with similar me aning [21, 56]. Yokoyama and Klyuev [57] used japans W ordNet to find synonyms for each query term. Early work has focused on using Thesaurus to find synonyms and relat ed words to a query term. One of the most well-known the sauri is WordNet. It is a global lexical database which orga

nizes the words into sets of synonyms called synsets, then makes semantic relationships between those synsets. If all the synonyms of a word are used, some inappropriate syno nyms are returned by WordNet [58]. To overcome this pro blem, synonym of a word that has the same POS are only used. Gong and Cheang [59] created TSN (Term Semantic Network) to rule out the generated expanded terms with lo wer supports and confidence to the query word. As long as term in query May falls into different synsets, a synset wit h a similar meaning to the query term is chosen by conside ring the adjacent query terms which can be best matched w ith the terms present in each Synset. After selecting the mo st relevant synset, all the synonyms of the query term in th e synset plus the terms contained in any synset directly rel ated to synset are selected as expansion terms. Using Word Net may not be effective due to its reliance on predefined l ists of senses and lack of the ability to include new words. Moreover, it does not take care about the context of query because it does not consider query as whole [11]. Hence, q uery expansion may not be completed when query term is not listed in WordNet. Furthermore, partial match is a criti cal problem where there is no exact match between query and concepts.

6.2 Based On domain specific thesauri (Domain ont ology):

Ontology is a conceptualization of implicit domain concept knowledge into explicit knowledge and human-machine understandable format [60]. Ontology transforms the implicit knowledge into explicit knowledge, allowing members to access and share the field knowledge. Ontologies Development is a costly and laborious process. Bhogal, MacFarlane and Smith [58] proposed utilize the hierarchy of ontology concepts and its relation to obtain candidate terms by calculating the similarity between query word and ontology concepts. To enhance the conceptual similar calculation, Alqadah and Bhatnagar [61] proposed use Formal Concept Analysis (FCA) as similar measurement, a data analysis theory has been applied in many fields of information studies such as: data mining and knowledge ontology. Conesa [80] expanded semantically user query terms by integrating the advantages of ontology and WordNet.

6.3 Word Embedding Based:

Traditional information retrieval (IR) models consider ter ms as atomic units of information, disregarding the semant ic commonalities and the complex syntactic relationships i nterweaving them in the discourse. One of the direct impli cations of this strong assumption is the vocabulary mismat ch.
To overcome these weaknesses, Semantic term representat ion may be learned implicitly by utilizing document co-oc currence statistics LSI [62], (PLSI) Probabilistic latent sem antic analysis [63], (LDA) [Latent Dirichlet allocation][64] although it is efficient, but it needs high amount of time to learn (high computational Costs), require high computatio nal requirements. To address this issue explicitly learning word embedding was suggested by Network Language Mo dels (NLM) [65].

Recently, deep learning has received a great deal of attenti ons in the field of natural language processing. Word embe dding also known as distributed representations is a set of l anguage modeling such as word2vec [66] and GloVe [67]. Both are based on distribution hypothesis – words with si milar meaning have similar representations. Word embedd ings represents the semantic and syntactic information of words and their context to vector of real number reflects si milarity and dissimilarities between terms. Let x and y are two terms and x, y are their embedding, then the distance b etween x, y represents the semantic relatedness between x and y. Word embeddings can be trained and used to derive similarities and relations between words. In general, Word embedding provides a global representation of terms whic h may be appropriate for a specific word within a global co ntext, while the meaning of terms can vary dramatically by topic and locally surrounding words (one word has one me aning per discourse). Hence particular embedding may be appropriate for a specific word. Diaz and Mitra [68] showe d that topic-specific representations outperform global repr esentations.

Word embedding has shown its power in natural language processing [69, 70 and 81] and in query expansion [68, 71] over traditional approach. Word2vec has been used to enha nce the accuracy of query expansion. It expands queries wi th contextually associated words generated from word emb edding instead of synonyms from external resources. Most terms similar to user query are selected by computing its c osine distance from the words in vector space. Recently, fu zzy theory has been used to enhance the performance of Q uery expansion. Liu [72] Used Fuzzy Rules to assign a wei ght which determines the similarity between expansion ter ms generated from word embedding and the original query terms. Lin and Wang [73] used fuzzy rules to infer the wei ghts of the additionally generated terms based on relevanc e feedback methods. Although word embeddings are one o f the few successful unsupervised learning approaches, it h

as some drawbacks. One of the most drawbacks is that wor d embeddings cannot build a vector for words which have not been encountered during training time. Furthermore, w ord2vec assigns different representations to the words whi ch have the same morphologically analysis such as: hopele ss, flawless, and careless.

7. Methods Combination (Hybrid Methods):

In order to increase the effectiveness of query expansion a pproach, basic methods are used in conjunction with other methods. As we described above, statistical method depen ds on term frequency to generate expansion features, but it does not consider the meaning or the term dependency. On contrast semantic method depends on knowledge base whi ch considers the context, but it suffers from the limited nu mber of terms and relations in it. The most promising meth od is a hybrid method which utilizes statistical method to g enerate expansion term and semantic method which guaran tees the valid order of terms in its correct context. In [74, 7 5] researcher combined between two methods in order to i ncrease the accuracy of query expansion. The main idea of this method is to tackle the Emperor interpretation of quer y in statistical method by adding words represented the me aning of the query.

As document length increases, there is a higher probability that the document loses its informative representation. To address this issue, Text summarization is an effective meth od which aims to find more informative representation of d ocument in a few sentences. Chang and Ounis [76] used th e summarized documents as a source for expansion. Altho ugh this method is effective, its quality depends on the effe ctiveness of methods which are used to summarize the doc ument. Song [75] showed query expansion method by com bining association role with external knowledge source suc h as (DBpedia).

Table I. query expansion method advantage and disadvantage

Method	*Advantage*	*Disadvantage*
Statistical approach (Term frequency)	- Easy to compute - Perform better than lexical approaches	- Time consuming - Does not care about semantically similar terms - Does not work well in short text
Statistical approach (Term Co-occurrence)	- Achieves good performance	- Co-occurrence between terms not necessary mean the term semantically related - The position of terms does not taken into account
lexical resources (WordNet)	- This approach gives better results if domain specific thesaurus or dictionaries are used	- Build the Resources take a lot of time to and are expensive. - Automatically developed resources are cheap and faster, but are inaccurate
Morphology analyzes (Linguistic Method)	- Improve recall - Reduce the dimensionality of text - Reduce inflectional forms of words to a common base which often communicate the same meaning.	- The stem of a term represents a broader concept than the original term (Over generalization problem). - Complexity of Breaking down words. - Decrease precision
Syntactic analysis (Linguistic Method)	- Successful approach - Syntax conveys meaning in most languages because order and dependency contribute to connotation	- Require exact match. - Assignment of roles (subject, object, etc.) to the different participants in a sentence is difficult task - Word may play as different parts of speech in different contexts - The structure of a sentence may have different several possible interpretations
Word embedding (Semantic Method)	- Representative meaning of the query Contextually rather than find synonyms from external resources	- In ability to handle unknown word - No Shared Representations at sub-word levels
Hybrid approach	- Achieved the best results more robust	- Time consuming

Table 2. Mean Average Precision of Several AQE Methods on Comparable Test Collections.

		Mean Average Precision(MAP)								
		Dataset								
Reference	Method	TREC 2001/2002	TREC 2001	TREC 2002	TREC 6	TREC-12	custom	robust	web	RCV1
[17]	LGD	0. 2894	0. 318	0. 274						
[17, 72]	BM25	0. 289	0. 332	0. 267						0. 408
[17]	BM25+ glove	0. 3157	0. 359	0. 2937						
[71]	Co-occurrence				0. 2634					
[50]	Syntactic analysis					0. 1749				
[72]	WordNet									0. 436
[57]	WordNet+Relavnce Feedback						0. 53			
[77]	WordNet +Ontology						0. 9697			
[68]	Glove(local)					0. 563		0. 517	0. 258	
[68]	Glove(Global)					0. 545		0. 472	0. 232	
[72]	word2vec(CBOW)									0. 437
[72]	word2vec (Skipgram)									0. 436

8. Conclusions

There are huge amount of data available on the web but we cannot use it efficiently and effectively due to word misma tch problem. Query Expansion approaches have been emer ged as most promising method to overcome this issue. The re are various techniques on query expansion (AQE) to ma ke those techniques better known and more widely used. T his article has explained query expansion (AQE) in more d etails, reported advantages and disadvantages of each appr oach and discussed its performance.

The study has gone some ways towards enhancing our und erstanding of query expansion types, assists in understandi ng the role of each one, and provides comparable views be tween them. The results of this study highlights that the pe rformance in term of precision and recall measures in tradi tional method showed underperform semantic method whe reas the performance in semantic method showed better pe rformance but it does not match the high performance in h ybrid method.

References

1. Ooi, J., et al. A survey of query expansion, query suggestion and query refinement techniques. in Software Engineering and Computer Systems (ICSECS), 2015 4th International Conference on. 2015. IEEE.
2. Xu, J. and W. B. Croft. Query expansion using local and global document analysis. in Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. 1996. ACM.
3. Seuss, D. Ten years into the web, and the search problem is nowhere near solved. in Computers In Libraries Conference, March. 2004.
4. Lau, T. and E. Horvitz, Patterns of search: analyzing and modeling web query refinement, in UM99 User Modeling. 1999, Springer. p. 119-128.
5. Jansen, B. J., D. L. Booth, and A. Spink, Determining the informational, navigational, and transactional intent of Web queries. Information Processing & Management, 2008. 44(3): p. 1251-1266.
6. Sadowski, C., K. T. Stolee, and S. Elbaum. How developers search for code: a case study. in Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering. 2015. ACM.
7. Carpineto, C. and G. Romano, A survey of automatic query expansion in information retrieval. ACM Computing Surveys (CSUR), 2012. 44(1): p. 1.
8. Furnas, G. W., et al., The vocabulary problem in human-system communication. Communications of the ACM, 1987. 30(11): p. 964-971.
9. Robertson, S. E. and K. S. Jones, Relevance weighting of search terms. Journal of the Association for Information Science and Technology, 1976. 27(3): p. 129-146.
10. Erritali, M., Information Retrieval: Textual Indexing Using an Oriented Object Database. Indonesian Journal of Electrical Engineering and Computer Science, 2016. 2(1): p. 205-214.
11. Manning, C. D., P. Raghavan, and H. Schütze, Introduction to information retrieval. Vol. 1. 2008: Cambridge university press Cambridge.
12. Luhn, H. P., The automatic creation of literature abstracts. IBM Journal of research and development, 1958. 2(2): p. 159-165.
13. Fang, H. and C. Zhai. An exploration of axiomatic approaches to information retrieval. in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. 2005. ACM.
14. Jung, Y., H. Park, and D. -z. Du, An effective term-weighting scheme for information retrieval. Computer Science Technical Report TR008, Department of Computer Science, University of Minnesota, Minneapolis, Minnesota, 2000: p. 1-15.
15. Efron, M., Linear time series models for term weighting in information retrieval. Journal of the Association for Information Science and Technology, 2010. 61(7): p. 1299-1312.
16. Cao, G., et al. Selecting good expansion terms for pseudo-relevance feedback. in Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. 2008. ACM.
17. El Mahdaouy, A., S. O. El Alaoui, and É. Gaussier. Semantically enhanced term frequency based on word embeddings for Arabic information retrieval. in Information Science and Technology (CiSt), 2016 4th IEEE International Colloquium on. 2016. IEEE.
18. Attar, R. and A. S. Fraenkel, Local feedback in full-text retrieval systems. Journal of the ACM (JACM), 1977. 24(3): p. 397-417.
19. Bai, J., et al. Query expansion using term relationships in language models for information retrieval. in Proceedings of the 14th ACM international conference on Information and knowledge management. 2005. ACM.
20. Bai, J., et al. Using query contexts in information retrieval. in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007. ACM.
21. Shaalan, K., S. Al-Sheikh, and F. Oroumchian. Query expansion based-on similarity of terms for improving Arabic information retrieval. in International Conference on Intelligent Information Processing. 2012. Springer.
22. Wei, J., S. Bressan, and B. C. Ooi. Mining term association rules for automatic global query expansion: methodology and preliminary results. in Web Information Systems Engineering, 2000. Proceedings of the First International Conference on. 2000. IEEE.
23. Lv, Y., C. Zhai, and W. Chen. A boosting approach to improving pseudo-relevance feedback. in Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011. ACM.

24. Lv, Y. and C. Zhai. Positional relevance model for pseudo-relevance feedback. in Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010. ACM.
25. Xu, Y., G. J. Jones, and B. Wang. Query dependent pseudo-relevance feedback based on wikipedia. in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. 2009. ACM.
26. Nie, L., et al., Query expansion based on crowd knowledge for code search. IEEE Transactions on Services Computing, 2016. 9(5): p. 771-783.
27. Zhai, C. and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. in Proceedings of the tenth international conference on Information and knowledge management. 2001. ACM.
28. Bade, Y., R. Bhat, and P. Borate, Optimization techniques for improving the performance of information retrieval system. International Journal of Research in Advent Technology, 2014. 2(2): p. 263-267.
29. Raman, K., et al. On improving pseudo-relevance feedback using pseudo-irrelevant documents. in European Conference on Information Retrieval. 2010. Springer.
30. Paik, J. H., D. Pal, and S. K. Parui, Incremental blind feedback: An effective approach to automatic query expansion. ACM Transactions on Asian Language Information Processing (TALIP), 2014. 13(3): p. 13.
31. Al-Shboul, B. and S. -H. Myaeng. Analyzing topic drift in query expansion for Information Retrieval from a large-scale patent DataBase. in Big Data and Smart Computing (BIGCOMP), 2014 International Conference on. 2014. IEEE.
32. Salton, G. and C. Buckley, Improving retrieval performance by relevance feedback in Journal of the American Society for Information Science. Volume, 1990. 14: p. 288-297.
33. Macdonald, C. and I. Ounis. Expertise drift and query expansion in expert search. in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. 2007. ACM.
34. Sakai, T., T. Manabe, and M. Koyama, Flexible pseudo-relevance feedback via selective sampling. ACM Transactions on Asian Language Information Processing (TALIP), 2005. 4(2): p. 111-135.
35. Saneifar, H., et al., Enhancing passage retrieval in log files by query expansion based on explicit and pseudo relevance feedback. Computers in Industry, 2014. 65(6): p. 937-951.
36. Rahman, M. M., S. K. Antani, and G. R. Thoma, A query expansion framework in image retrieval domain based on local and global analysis. Information processing & management, 2011. 47(5): p. 676-691.
37. Ko, Y. and J. Seo, Text classification from unlabeled documents with bootstrapping and feature projection techniques. Information Processing & Management, 2009. 45(1): p. 70-83.
38. Kelly, D. and J. Teevan. Implicit feedback for inferring user preference: a bibliography. in Acm Sigir Forum. 2003. ACM.
39. Okabe, M. and S. Yamada, Semisupervised query expansion with minimal feedback. IEEE Transactions on Knowledge and Data Engineering, 2007. 19(11).
40. Diaz, F. and D. Metzler. Improving the estimation of relevance models using large external corpora. in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006. ACM.
41. Weerkamp, W., K. Balog, and M. de Rijke, Exploiting external collections for query expansion. ACM Transactions on the Web (TWEB), 2012. 6(4): p. 18.
42. Liu, Z., S. Natarajan, and Y. Chen, Query expansion based on clustered results. Proceedings of the VLDB Endowment, 2011. 4(6): p. 350-361.
43. Oh, H. -S. and Y. Jung, Cluster-based query expansion using external collections in medical information retrieval. Journal of biomedical informatics, 2015. 58: p. 70-79.
44. Harper, D. J. and C. J. Van Rijsbergen, An evaluation of feedback in document retrieval using co-occurrence data. Journal of documentation, 1978. 34(3): p. 189-216.
45. Kadri, Y. and J. -Y. Nie. Effective stemming for Arabic information retrieval. in proceedings of the Challenge of Arabic for NLP/MT Conference, Londres, Royaume-Uni. 2006.
46. Boudchiche, M., et al., AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer. Journal of King Saud University-Computer and Information Sciences, 2017. 29(2): p. 141-146.
47. Pasha, A., et al. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. in LREC. 2014.
48. Al-Serhan, H. and A. Ayesh. A triliteral word roots extraction using neural network for Arabic. in Computer Engineering and Systems, The 2006 International Conference on. 2006. IEEE.
49. Elayeb, B. and I. Bounhas, Arabic cross-language information retrieval: a review. ACM Transactions on Asian and Low-Resource Language Information Processing, 2016. 15(3): p. 18.
50. Sun, R., C. -H. Ong, and T. -S. Chua. Mining dependency relations for query expansion in passage retrieval. in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006. ACM.
51. Punyakanok, V., D. Roth, and W. -t. Yih, The importance of syntactic parsing and inference in semantic role labeling. Computational Linguistics, 2008. 34(2): p. 257-287.
52. Selvaretnam, B. and M. Belkhatir, A linguistically driven framework for query expansion via grammatical constituent highlighting and role-based concept weighting. Information Processing & Management, 2016. 52(2): p. 174-192.
53. Momtazi, S., S. Khudanpur, and D. Klakow. A comparative study of word co-occurrence for term clustering in language model-based sentence retrieval. in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010. Association for Computational Linguistics.
54. Castells, P., M. Fernandez, and D. Vallet, An adaptation of the vector-space model for ontology-based information retrieval. IEEE transactions on knowledge and data engineering, 2007. 19(2).
55. Kassim, J. M. and M. Rahmany. Introduction to semantic search engine. in Electrical Engineering and Informatics, 2009. ICEEI'09. International Conference on. 2009. IEEE.
56. Ray, S. K., S. Singh, and B. P. Joshi, A semantic approach for question classification using WordNet and Wikipedia. Pattern Recognition Letters, 2010. 31(13): p. 1935-1943.

57. Yokoyama, A. and V. Klyuev. Search Engine Query Expansion Using Japanese WordNet. in Human-Centric Computing (HumanCom), 2010 3rd International Conference on. 2010. IEEE.
58. Bhogal, J., A. MacFarlane, and P. Smith, A review of ontology based query expansion. Information processing & management, 2007. 43(4): p. 866-886.
59. Gong, Z., C. W. Cheang, and U. L. Hou. Web query expansion by WordNet. in International Conference on Database and Expert Systems Applications. 2005. Springer.
60. Fensel, D., et al., OIL: An ontology infrastructure for the semantic web. IEEE intelligent systems, 2001. 16(2): p. 38-45.
61. Alqadah, F. and R. Bhatnagar, Similarity measures in formal concept analysis. Annals of Mathematics and Artificial Intelligence, 2011. 61(3): p. 245-256.
62. Deerwester, S., et al., Indexing by latent semantic analysis. Journal of the American society for information science, 1990. 41(6): p. 391.
63. Hofmann, T. Probabilistic latent semantic analysis. in Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. 1999. Morgan Kaufmann Publishers Inc.
64. Blei, D. M., A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation. Journal of machine Learning research, 2003. 3(Jan): p. 993-1022.
65. Sergienko, R., et al., Collectives of Term Weighting Methods for Natural Language Call Routing, in Informatics in Control, Automation and Robotics. 2016, Springer. p. 99-110.
66. Mikolov, T., et al. Distributed representations of words and phrases and their compositionality. in Advances in neural information processing systems. 2013.
67. Pennington, J., R. Socher, and C. Manning. Glove: Global vectors for word representation. in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
68. Diaz, F., B. Mitra, and N. Craswell, Query expansion with locally-trained word embeddings. arXiv preprint arXiv:1605. 07891, 2016.
69. Zhou, G., et al. Learning continuous word embedding with metadata for question retrieval in community question answering. in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015.
70. Zhang, M., et al. Building Earth Mover's Distance on Bilingual Word Embeddings for Machine Translation. in AAAI. 2016.
71. Roy, D., et al., Using word embeddings for automatic query expansion. arXiv preprint arXiv:1606. 07608, 2016.
72. Liu, Q., et al. Enhanced word embedding similarity measures using fuzzy rules for query expansion. in Fuzzy Systems (FUZZ-IEEE), 2017 IEEE International Conference on. 2017. IEEE.
73. Lin, H. -C., L. -H. Wang, and S. -M. Chen, Query expansion for document retrieval based on fuzzy rules and user relevance feedback techniques. Expert Systems with Applications, 2006. 31(2): p. 397-405.
74. Bai, J., J. -Y. Nie, and G. Cao. Context-dependent term relations for information retrieval. in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. 2006. Association for Computational Linguistics.
75. Song, M., et al., Integration of association rules and ontologies for semantic query expansion. Data & Knowledge Engineering, 2007. 63(1): p. 63-75.
76. Chang, Y., I. Ounis, and M. Kim, Query reformulation using automatically generated query concepts from a document space. Information processing & management, 2006. 42(2): p. 453-468.
77. Chauhan, R., et al. Domain ontology based semantic search for efficient information retrieval through automatic query expansion. in Intelligent Systems and Signal Processing (ISSP), 2013 International Conference on. 2013. IEEE.
78. pink, A., et al., Searching the web: The public and their queries. Journal of the Association for Information Science and Technology, 2001. 52(3): p. 226-234.
79. Chinnakotla, M. K. (2010). Information Retrieval in Multilingual ResourceConstrained Settings. PhD thesis, Indian Institute of Technology, Mumbai.
80. Conesa, J., V.C. Storey, and V. Sugumaran, Improving web-query processing through semantic knowledge. Data & Knowledge Engineering, 2008. 66(1): p. 18-34.
81. Almasri, M., C. Berrut, and J.-P. Chevallet. A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information. in European conference on information retrieval. 2016. Springer.
82. Gao, L., et al. Query expansion for exploratory search with subtopic discovery in Community Question Answering. in Neural Networks (IJCNN), 2016 International Joint Conference on. 2016. IEEE.