A New Approach for Query Expansion using Wikipedia and WordNet

Preprint · January 2019

2 authors:

Hiteshwar Kumar Azad
K L University
10 PUBLICATIONS **90** CITATIONS

SEE PROFILE

Akshay Deepak
National Institute of Technology Patna
41 PUBLICATIONS **111** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Query Expansion View project

Modelling Protein Sequences View project

A New Approach for Query Expansion using Wikipedia and WordNet

Hiteshwar Kumar Azad · Akshay Deepak

Abstract Query expansion (QE) is a well known technique to enhance the effectiveness of information retrieval (IR). QE reformulates the initial query by adding similar terms that helps in retrieving more relevant results. Several approaches have been proposed with remarkable outcome, but they are not evenly favorable for all types of queries. One of the main reasons for this is the use of the same data source while expanding both the individual and the phrase query terms. As a result, the holistic relationship among the query terms is not well captured. To address this issue, we have selected separate data sources for individual and phrase terms. Specifically, we have used WordNet for expanding individual terms and Wikipedia for expanding phrase terms. We have also proposed novel schemes for weighting expanded terms: *in-link score* (for terms extracted from Wikipedia) and a tfidf based scheme (for terms extracted from WordNet). In the proposed Wikipedia-WordNet based QE technique (WWQE), we weigh the expansion terms twice: first, they are scored by the weighting scheme individually, and then, the weighting scheme scores the selected expansion terms in relation to the entire query using correlation score. The experimental results show that the proposed approach successfully combines Wikipedia and WordNet as demonstrated through a better performance on standard evaluation metrics on FIRE dataset. The proposed WWQE approach is also suitable with other standard weighting models for improving the effectiveness of IR.

Keywords Query Expansion · Information Retrieval · WordNet · Wikipedia

1 Introduction

Web is arguably the largest information source available on this planet and it's growing day by day. According to a recent survey [26] of the computer world magazine, approximately 70-80 percent of all data available to enterprises/organizations is unstructured information, i.e., information that either does not organize in a pre-defined manner or does not have a pre-defined data model. This makes information processing a big challenge and, creates a vocabulary gap between user queries and indexed documents. It is common for a user's query Q and its relevant document D (in a document collection) to use different vocabulary and language styles while referring to the same concept. For example, terms 'buy' and 'purchase' have the same meaning, only one of these can be present in documents-index while the other one can be user's query term. This makes it difficult to retrieve the information actually wanted by the user. An effective strategy to fill this gap is to use Query expansion (QE) technique that enhances the retrieval effectiveness by adding expansion terms to the initial query. Selection of the expansion terms plays a crucial role in QE because only a small subset of the expanded terms are actually relevant to the query. In this sense, the

———————————
Hiteshwar Kumar Azad
National Institute of Technology Patna
E-mail: hiteshwar.cse15@nitp.ac.in

Akshay Deepak
National Institute of Technology Patna

approach for selection of expansion terms is equally important in comparison to what we do further with the expanded terms in order to retrieve desired information. QE has a long research history in Information retrieval (IR) [64, 83]. It has potential to enhance the IR effectiveness by adding relevant terms that can help to discriminate the relevant documents from irrelevant ones. The source of expansion terms plays a significant role in QE. A variety of sources have been researched for extracting the expansion terms, e.g., the entire target document collection [14, 24, 110], feedback documents (few top ranked documents are retrieved in response to the initial query) [31, 59] or external knowledge resources [1, 33, 54].

References [10, 25] provide comprehensive surveys on data sources used for QE. Broadly, such sources can be classified into four categories: documents used in retrieval process [14] (e.g., corpus), hand-built knowledge resources [76] (e.g., WordNet[1], ConceptNet[2], thesaurus, ontologies), external text collections and resources [1] (e.g., Web, Wikipedia), and hybrid data sources [32].

In corpus based sources, a corpus is prepared that contains a cluster of terms for each possible query term. During expansion, the corresponding cluster is used as the set of expanded terms (e.g., [14, 24, 110]). However, corpus based sources fail to establish a relationship between a word in the corpus and related words used in different communities, e.g., "senior citizen" and "elderly" [39].

Hand-built knowledge resources based QE extract knowledge from textual hand-built data sources such as dictionaries, thesaurus, ontologies and LOD cloud (e.g., [9, 76, 95, 102, 108]). Thesaurus-based QE can be either automatic or hand-built. One of the famous hand-built thesaurus is Word-Net [66]. While it significantly improves the retrieval effectiveness of badly constructed queries, it does not show a lot improvement for well formulated user queries. Primarily, there are three limitations of hand-built knowledge resources: they are commonly domain specific, they usually do not contain proper noun and they have to be kept up to date.

External text collections and resources such as web, Wikipedia, Query logs and anchor texts are the most common and effective data sources for QE ([1, 11, 16, 33, 54, 97, 106]). In such cases, QE approaches show overall better results in comparison to the other previously discussed data sources.

Hybrid Data Sources are a combination of two or more data sources. For example, reference [28] uses WordNet, an external corpus, and the top retrieved documents as data sources for QE. Some of the other research works based on hybrid resources are [32, 44, 57, 100].

Among the above data sources, Wikipedia and WordNet are popular choices for semantic enrichment of the initial query [1, 4, 38, 76, 95, 104]. They are also two of the most widely used knowledge resources in natural language processing. Wikipedia is the largest encyclopedia describing entities [99]. WordNet is a large lexicon database of words in the English language. An entity is described by Wikipedia through a web-article that contains detailed related information about the entity. Each such web-article describes only one entity. The information present in the article has important keywords that can prove very useful as expansion terms for queries based on the entity being described by the article. On the other hand, WordNet consists of a graph of synsets that are collections of synonymous words linked by a number of useful properties. WordNet also provides a precise and attentively assembled hierarchy of useful concepts. These features make WordNet an ideal knowledge resource for QE.

Many of the articles [1, 4, 38, 60, 76, 95, 104] have used Wikipedia and WordNet separately with promising results. However, they don't produce consistent results for different types of queries (individual and phrase queries).

This article proposes a novel technique named *Wikipedia-WordNet based QE technique* (WWQE) for query expansion that combines Wikipedia and WordNet data sources to improve retrieval effectiveness. We have also proposed novel schemes for weighting expanded terms: *in-link score* (for terms extracted from Wikipedia) and a tfidf based scheme (for terms extracted from WordNet). Experimental results show that the proposed WWQE technique produces consistently better results for all kinds of queries (individual and phrase queries) when compared with query expansion based on the two data sources individually. The experiments were carried on FIRE dataset [50] using popular weighting models and evaluation matrices.

---

[1] https://wordnet.princeton.edu/

[2] http://conceptnet5.media.mit.edu/

1.1 Contributions

The following are the contributions of this paper:

− *Data Sources.* A novel technique for query expansion named *Wikipedia-WordNet based QE technique* (WWQE) is proposed that combines Wikipedia and WordNet as data sources. To the best of our knowledge, these two data sources have not been used together for QE.
− *Term Selection.* Proposed WWQE technique employs a two level strategy to select terms from WordNet. For fetching expansion terms from Wikipedia pages of the query terms, the proposed technique uses a novel weighting scheme based on out-links and in-links called *in-link score.*
− *Phrase Selection.* To compute similarity between query terms and a document, the proposed WWQE technique employs separate measures for phrases and non-phrase query terms.
− *Weighting Method.* For weighting candidate expansion terms obtained from Wikipedia, the proposed WWQE technique uses a novel weighting scheme based on out-links and in-links, and correlation score. For terms obtained from WordNet, it uses a novel tf-idf and correlation score based weighting scheme.
− Experiments were conducted on Forum for Information Retrieval Evaluation (FIRE) collections. They produced improved results on popular metrics such as MAP(mean average precision), GM_MAP (geometric mean average precision), P@10 (precision at top 10 ranks), P@20, P@30, bpref (binary preference) and overall recall. The comparison was made with results obtained on individual data sources (i.e., Wikipedia and WordNet).

1.2 Organization

The remainder of the article is organized as follows. Section 2 discusses related work. Section 3 describes the proposed approach. Experimental Setup, dataset and evaluation matrices are discussed in Section 4. Section 5 discusses the experimental results. Finally, we conclude in Section 6.

2 Related Work

Query Expansion has rich literature in the area of Information Retrieval (IR). In the era of 1960s, Moron et al. [64] was the first researcher who applied QE for literature indexing and searching in a mechanized library system. In 1971, Rocchio [83] brought QE to spotlight through "relevance feedback method" and its characterization in a vector space model. This method is still used in its original and modified forms in automatic query expansion (AQE). Rocchio's work was further extended and applied in techniques such as collection-based term co-occurrence [52, 81], cluster-based information retrieval [51, 68], comparative analysis of term distribution [78, 93, 107] and automatic text processing [84–86].

Recently, QE has come to spotlight because a lot of researchers are using QE techniques for working on personalized social bookmarking services [18, 21, 40], Question Answering over Linked Data (QALD)[3] [92], Text Retrieval Conference (TREC)[4] and Forum for Information Retrieval Evaluation (FIRE)[5] collections . They are also used heavily in web, desktop and email searches [75]. Many platforms provide QE facility to end users, which can be turned on or off, e.g., WordNet[6], ConceptNet[7], Lucene[8], Google Enterprise [9] and MySQL. Some surveys have previously been done on QE techniques. In 2007, Bhogal et al. [17] reviewed QE techniques using ontologies, which are domain specific. Such techniques have also been described in book [63]. Carpineto et al. [25] reviewed major QE techniques, Data sources and features in an information retrieval system. In this paper we propose an AQE technique based on WordNet and Wikipedia, which are currently highly influential data sources. These two sources are described next.

---

[3] http://qald.sebastianwalter.org/

[4] http://trec.nist.gov/

[5] http://fire.irsi.res.in/

[6] https://wordnet.princeton.edu/

[7] http://conceptnet5.media.mit.edu/

[8] http://lucene.apache.org/

[9] https://enterprise.google.com/search/

Use of WordNet as Data Source for QE

WordNet [66] is one of the popular hand-built thesaurus, which has been significantly used for QE and word-sense disambiguation (WSD). Here, our focus is on the use of WordNet for query expansion. There are many issues that need to be addressed when using WordNet as a data source, such as:

- When a query term appears in multiple synsets, which synset(s) should be considered for query expansion!
- Can only the synsets of a query term have meanings similar to the query term, or, synsets of these synsets can also have meanings similar to the query term, and hence, should also be considered as potential expansion terms!
- When considering a synset of a query term, should only synonyms be considered or other relations (i.e., hypernyms, hyponyms, holonyms, meronyms etc.) should also be looked at! Further, when considering terms under a given relation, which terms should be selected!

In earlier works, a number of researchers have explored these issues. References [94, 95] added manually selected WordNet synsets for QE, but unfortunately no significant improvement were obtained. Reference [87] uses synonyms of the initial query and assigns half weight. Reference [60] used word sense to add synonyms, hyponyms and terms's WordNet glosses to expand query. Their experiments yielded significant improvements on TREC datasets. Reference [41] uses semantic similarity while reference [108] uses sense disambiguation of query terms to add synonyms for QE. During experimental evaluation, in response to the user's initial query, reference [108]'s method produces an improvement of around 7% in P@10 value over the CACM collection. Reference [35] uses a set of candidate expansion terms (CET) that include all the terms from all the synsets where the query terms exist. Basically, a CET is chosen based on the vocabulary overlap between its glosses and the glosses of query terms.

Recently, reference [76] used semantic relations from the WordNet. The authors proposed a novel query expansion technique where Candidate Expansion Terms (CET) are selected from a set of pseudo-relevant documents. The usefulness of these terms is determined by considering multiple sources of information. The semantic relation between the expanded terms and the query terms is determined using WordNet. On the TREC collection, their method showed significant improvement in IR over the user's unexpanded queries. Reference [58] presents an automatic query expansion (AQE) approach that uses word relations to increase the chances of finding relevant code. As data source for query expansion, it uses a thesaurus containing only software-related word relations along with WordNet. More recently, reference [62] used WordNet for effective code search, where it was used to generate synonyms, which were used as query expansion terms. During experimental evaluation, their approach showed improvement in precision and recall by values by 5% and 8% respectively.

In almost all the aforementioned studies, CETs are taken from WordNet as synsets of initial queries. In contrast, we selected CETs from not only the synsets of the initial query, but also synsets of these synsets. We then assign weights to the synonyms level wise.

Use of Wikipedia as Data Source for QE

Wikipedia [99] is a freely available and the largest multilingual Online encyclopedia on the web, where articles are regularly updated and new articles are added by a large number of web users. The exponential growth and reliability of Wikipedia makes it an ideal knowledge resource for information retrieval.

Recently, Wikipedia is being used widely for QE and a number of studies have reported significant improvements in IR over TREC and Cultural Heritage in CLEF (CHiC) datasets (e.g., [1,4,7,34,43,59,104]). Reference [59] performed an investigation using Wikipedia and retrieved all articles corresponding to the original query as a source of expansion terms for pseudo relevance feedback. It observed that for a particular query where the usual pseudo relevance feedback fails to improve the query, Wikipedia-based pseudo relevance feedback improves it significantly. Reference [34] uses link-based QE on Wikipedia and focuses on anchor text. It also proposed a phrase scoring function. Reference [104] utilized Wikipedia to categorize the original query into three

types: (1) ambiguous queries (queries with terms having more than one potential meaning), (2) entity queries (queries having a specific meaning that cover a narrow topic) and (3) broader queries (queries having neither ambiguous nor specific meaning). They consolidated the expansion terms into the original query and evaluated these techniques using language modeling IR. Reference [4] uses Wikipedia for semantic enrichment of short queries based on in-link and out-link articles. Reference [32] proposed Entity Query Feature Expansion (EQFE) technique. It uses data sources such as Wikipedia and Freebase to expand the initial query with features from entities and their links to knowledge bases (Wikipedia and Freebase). It also uses structured attributes and the text of the knowledge bases for query expansion. The main motive for linking entities to knowledge bases is to improve the understanding and representation of text documents and queries.

Our proposed WWQE method differs from the above mentioned expansion methods in two ways:

1. Our method uses both Wikipedia and WordNet for query expansion, whereas the above discussed methods either use only one of these sources or some other sources.
2. For extracting expansion terms from WordNet, our method employs a novel two level approach where synsets of the query term as well as the synsets of these synsets are selected.
3. For extracting expansion terms from Wikipedia, terms are selected on the basis of a novel scheme called 'in-link score', which is based on in-links and out-links of Wikipedia articles.

Other QE Approaches

On the basis of data sources used in QE, several approaches have been proposed. All these approaches can be classified into four main categories:

***Linguistic approaches*:** The approaches in this category analyze expansion features such as lexical, morphological, semantic and syntactic term relationships to reformulate the initial query terms. They use thesaurus, dictionaries, ontologies, Linked Open Data (LOD) cloud or other similar knowledge resources such as WordNet or ConceptNet to determine the expansion terms by dealing with each keyword of initial query independently.

Word stemming is one of the first and among the most influential QE approaches in linguistic association to reduce the inflected word to its root word. The stemming algorithm (e.g., [77]) can be utilized either at retrieval time or at indexing time. When used during retrieval, terms from initially retrieved documents are picked, and then, these terms are harmonized with the morphological types of query terms (e.g., [55, 73]). When used during indexing time, words picked from the document collection are stemmed, and then, these words are harmonized with the query root word stems (e.g., [49]). Morphological approach [55, 73] is an ordered way of studying the internal structure of the word. It has been shown to give better results than the stemming approach [20, 69], however, it requires querying to be done in a structured way.

Use of semantic and contextual analysis are other popular QE approaches in linguistic association. It includes knowledge sources such as Ontologies, LOD cloud, dictionaries and thesaurus. In the context of ontological based QE, reference [17] uses domain-specific and domain-independent ontologies. Reference [101] utilizes the rich semantics of domain ontology and evaluates the trade off between the improvement in retrieval effectiveness and the computational cost. Several research works have been done on QE using a thesaurus. WordNet is a well known thesaurus for expanding the initial query using word synsets. As discussed earlier, many of the research works use WordNet for expanding the initial query. For example, reference [95] uses WordNet to find the synonyms. Reference [87] uses WordNet and POS tagger for expanding the initial query. However, this approach suffers some practical issues such as absence of accurate matching between query and senses, absence of proper nouns, and, one query term mapping to many noun synsets and collections. Generally, utilization of WordNet for QE is beneficial only if the query words are unambiguous in nature [42, 95]; using word sense disambiguation (WSD) to remove ambiguity is not easy [71, 74]. Several research works have attempted to address the WSD problem. For example, reference [72] suggests that instead of considering the replacement of the initial query term with its synonyms, hyponyms, and hyperonyms, it is better to extract similar concepts from the same domain of the given query from WordNet (such as the common nodes and glossy terms).

Another important approach that improves the linguistic information of the initial query is syntactic analysis [109]. Syntactic based QE uses the enhanced relational features of the query terms for expanding the initial query. It expands the query mostly through statistical approaches [101].

It recognizes the term dependency statistically [80] by employing techniques such as term co-occurrence. Reference [89] uses this approach for extracting contextual terms and relations from external corpus. Here, it uses two dependency relation based query expansion techniques for passage retrieval: Density based system (DBS) and Relation based system (RBS). DBS makes use of relation analysis to extract high quality contextual terms. RBS extracts relation paths for QE in a density and relation based passage retrieval framework. The syntactic analysis approach may be beneficial for natural language queries in search tasks, where linguistic analysis can break the task into a sequence of decisions [109] or integrate the taxonomic information effectively [61].

However, the above approaches fail to solve ambiguity problems [10, 25].

Corpus-based approaches: Corpus-based Approaches examine the contents of the whole text corpus to recognize the expansion features to be utilized for QE. They are one of the earliest statistical approaches for QE. They create co-relations between terms based on co-occurrence statistics in the corpus to form sentences, paragraphs or neighboring words, which are used in the expanded query. Corpus-based approaches have two admissible strategies: (1) term clustering [29, 52, 68], which groups document terms into clusters based on their co-occurrences, and, (2) concept based terms [37, 70, 79], where expansion terms are based on the concept of query rather than the original query terms. Reference [56] selects the expansion terms after the analysis of the corpus using word embeddings, where each term in the corpus is characterized with a vector embedded in a vector space. Reference [110] uses four corpora as data sources (including one industry and three academic corpora) and presents a Two-stage Feature Selection (TFS) framework for QE known as Supervised Query Expansion (SQE).

Some of the other approaches established an association thesaurus based on the whole corpus by using, e.g., context vectors [39], term co-occurrence [24], mutual information [45] and interlinked Wikipedia articles [67].

Search log-based approaches: These approaches are based on the analysis of search logs. User feedback, which is an important source for suggesting a set of similar terms based on the user's initial query, is generally explored through the analysis of search logs. With the fast growing size of the web and the increasing use of web search engines, the abundance of search logs and their ease of use have made them an important source for QE. It usually contains user queries corresponding to the URLs of Web pages. Reference [30] uses the query logs to extract probabilistic correlations between query terms and document terms. These correlations are further used for expanding the user's initial query. Similarly, reference [31] uses search logs for QE; their experiments show better results when compared with QE based on pseudo relevance feedback. One of the advantages of using search logs is that it implicitly incorporates relevance feedback. On the other hand, it has been shown in reference [98] that implicit measurements are relatively good, however, their performance may not be the same for all types of users and search tasks.

There are commonly two types of QE approaches used on the basis of web search logs. The first type considers queries as documents and extracts features of these queries that are related to the user's initial query [47]. Among the techniques based on the first approach, some use their combined retrieval results [48], while some do not (e.g., [47, 106]).

In the second type of approach, the features are extracted on relational behavior of queries. For example, reference [12] represents queries in a graph based vector space model (query-click bipartite graph) and analyzes the graph constructed using the query logs. References [23, 31, 80] extract the expansion terms directly from the clicked results. References [36, 96] use the top results from past query terms entered by the users. Queries are also extracted from related documents [19, 97], or through user clicks [46, 105, 106]. The second type of approach is more popular and has been shown to give better results.

Web-based approaches: These approaches include Wikipedia and anchor texts from websites for expanding the user's original query. These approaches have gained popularity in recent times. Anchor text was first used in reference [65] for associating hyper-links with linked pages and with the pages in which anchor texts are found. In the context of a web-page, an anchor text can play a role similar to the title since the anchor text pointing to a page can serve as a concise summary of its contents. It has been shown that user search queries and anchor texts are very similar because an anchor text is a brief characterization of its target page. Article [54] used anchor texts for QE; their experimental results suggest that anchor texts can be used to improve the traditional QE based on query logs. On similar lines, reference [33] suggested that anchor texts can be an effective

substitute for query logs. It demonstrated effectiveness of QE techniques using log-based stemming through experiments on standard TREC collection dataset.

Another popular approach is the use of Wikipedia articles, titles and hyper-links (in-link and out-link) [4, 7]. We have already mentioned the importance of Wikipedia as an ideal knowledge source for QE. Recently, quite a few research works have used it for QE (e.g., [1, 4, 7, 59, 104]). Article [3] attempts to enrich initial queries using semantic annotations in Wikipedia articles combined with phrase-disambiguation. Their experiments show better results in comparison to the relevance based language model.

FAQs are another important web-based source of information for improving QE. Recently published article [53] uses domain specific FAQs data for manual QE. Some of the other works using FAQs are [2, 80, 88].

3 Our Approach

The proposed approach consist of four main steps: Pre-processing of Initial Query, QE using Wikipedia, QE using WordNet, and Re-weighting Expanded Terms. Figure 1 summarizes these steps.
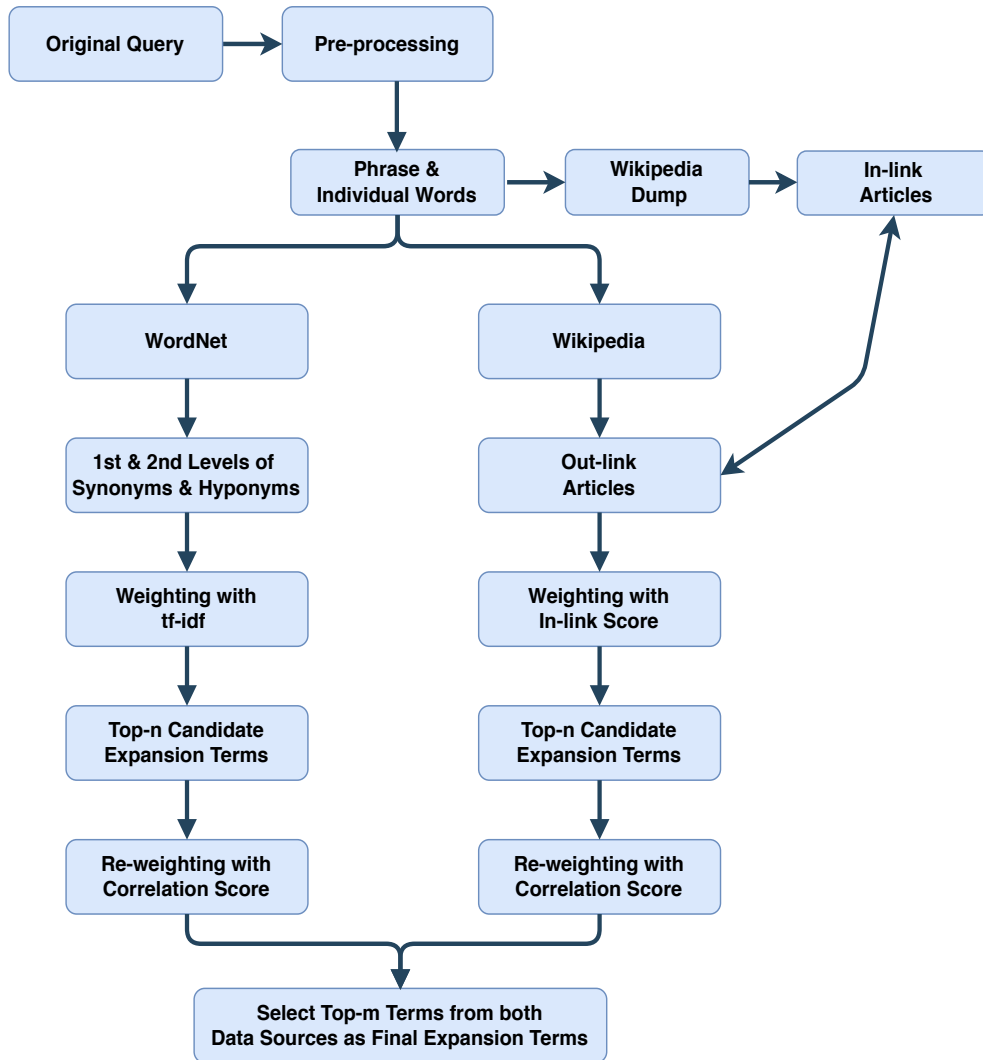

Fig. 1: Steps involved in the proposed approach

3.1 Pre-processing of Initial Query

In the Preprocessing step, Brill's tagger [22] is used to lematize each query and assign a Part of speech (POS) to each word in the query. The POS tagging is done on queries and the POS information is used to recognize the phrase and individual words. These phrases and individual words are used in the subsequent steps of QE. Many researchers agree that instead of considering the term-to-term relationship, dealing with the query in terms of phrases gives better results [3,31,61]. Phrases usually offer richer context and have less ambiguity. Hence, documents retrieved in response to phrases from the initial query have more importance than the documents retrieved in response to non-phrase words from the initial query. A phrase usually has a specific meaning that goes beyond the cumulative meaning of the individual component words. Therefor, we give more priority to phrases in the query than the individual words when finding expansion terms from Wikipedia and WordNet.

For example, consider the following query (Query ID- 126) from FIRE dataset to demonstrate our pre-processing approach:

$< top >$
$< num > 126 < /num >$
$< title >$Swine flu vaccine$< /title >$
$< desc >$Indigenous vaccine made in India for swine flu prevention$< /desc >$
$< narr >$Relevant documents should contain information related to making indigenous swine flu vaccines in India, the vaccine's use on humans and animals, arrangements that are in place to prevent scarcity / unavailability of the vaccine, and the vaccine's role in saving lives.$< /narr >$
$< /top >$

Multiple such queries in the standard SGML format are present in the query file of FIRE dataset. For extracting the root query, we extract the title from each query and tag it using the Stanford POS tagger library [91]. Fo example, the result of POS tagging the title of the above query is:

Swine_NN flu_NN vaccine_NN.

For extracting phrases, we have only considered Nouns, Adjectives and Verbs as the words of interest. We consider a phrase to have been identified whenever two or more consecutive Noun, Adjective or Verb words are found. Based on this, we get the following individual terms and phrases from the above query:

Swine
flu
Swine flu
vaccine
flu vaccine
Swine flu vaccine

3.2 QE using Wikipedia

After Pre-processing of the initial query we consider individual words and phrases as keywords to expand the initial query using Wikipedia. To select CETs from Wikipedia, we mainly focus on Wikipedia titles, in-links and out-links. Before going into further details, we first discuss our Wikipedia representation.

Wikipedia Representation
Wikipedia is an ideal information source for QE and can be represented as directed graph $G(A, L)$, where A and L indicate articles and links respectively. Each article $x \in A$ effectively summarizes its entity ($title(x)$) and provides links to the user to browse other related articles. In our work, we consider the two types of links: in-links and out-links.
In-links (I(x)): Set of articles that point to the article x. It can be defined as

$$I(x) = \{x_i \mid (x_i, x) \in L\} \tag{1}$$

For example, assume we have an article titled "Computer Science". The in-links to this article will be all the titles in Wikipedia that hyperlink to the article titled "Computer Science" in their main

text or body.

Out-links (O(x)): Set of articles that x point to. It can be defined as

$$O(x) = \{x_i \mid (x, x_i) \in L\} \tag{2}$$

For example, again consider the article titled "Computer Science". The out-links refers to all the hyperlinks within the body of the Wikipedia page of the article titled "Computer Science" (i.e. $https : //en.wikipedia.org/wiki/Computer\_Science$). The in-links and out-links have been diagrammatically demonstrated in Fig. 2.

Fig. 2: In-links & Out-links structure of Wikipedia

In addition to the article pages, Wikipedia contains "redirect" pages that provide an alternative way to reach the target article for abbreviated query terms. For example, query "ISRO" redirects to the article "Indian Space Research Organisation" and "UK" redirects to "United Kingdom".

In our proposed WWQE approach, the following steps are taken for QE using Wikipedia.

– *Extraction of In-links.*
– *Extraction of Out-links.*
– *Assignment of the in-link score to expansion terms.*
– *Selection top n terms as expansion terms.*
– *Re-weighting of expansion terms.*

Extraction of In-links

This step involves two sub-steps. First, extraction of In-links and, the second, computation of term frequency (tf) of the initial query terms. The in-links of an initial query term consist of titles of all those Wikipedia articles that contain a hyper-link to the given query term in their main text or body. tf of an initial query term is the term frequency of the initial query term and its synonyms obtained from WordNet in the in-link articles (see the Fig. 3). For example, if the initial query term is "Bird", and "Wings" is one of its in-links, then tf of "Bird" in the article "Wings" is the frequency of word "Bird" and its synonyms obtained from WordNet in the article "Wings".
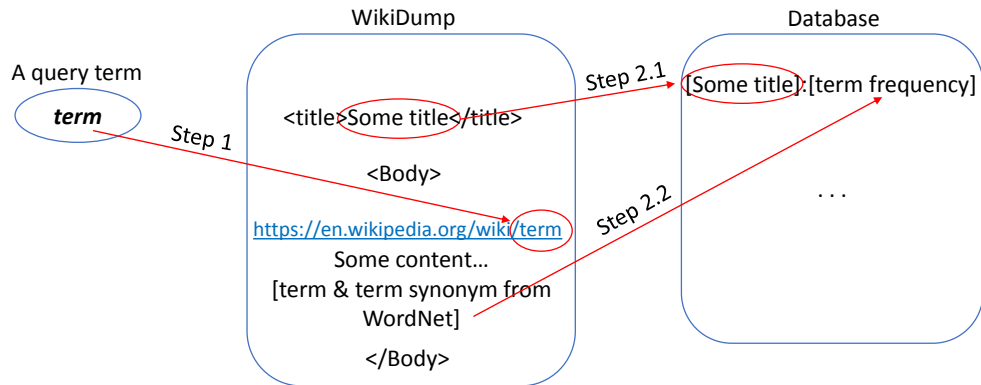

Fig. 3: In-links Extraction

Extraction of Out-links

Out-links of a query term are extracted by extracting the hyperlinks from the Wikipedia page of the query term as shown in Fig. 4. For example, if the initial query term is "Bird" then all the hyper-links within the body of the article "Bird" are extracted as out-links.
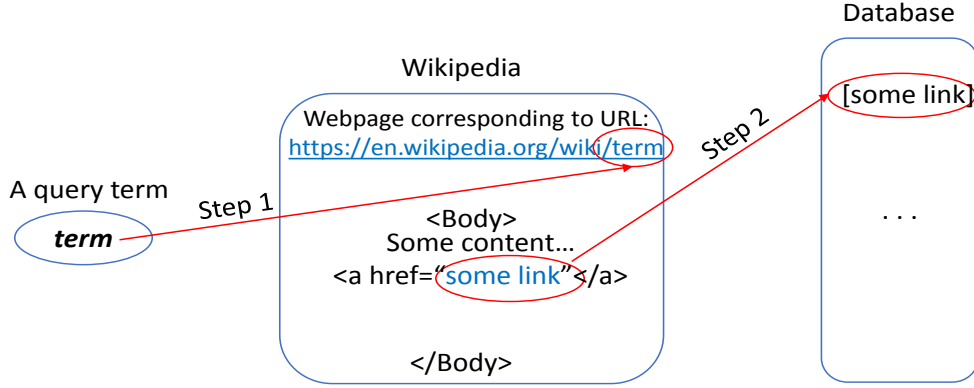

Fig. 4: Out-links Extraction

Assigning in-link score to expansion terms

After extraction of in-links and out-links of the query term, expansion terms are selected from the out-links on the basis of semantic similarity. Semantic similarity has been calculated based on in-link scores. Let t be a query term and t_1 be its one of the candidate expansion terms. In reference to Wikipedia, these two articles t and t_1 are considered to be semantically similar if (i) t_1 is both an out-link and an in-link of t and (ii) t_1 has a high in-link score. The in-link score is based on the popular weighting scheme $tf \cdot idf$ [86] in IR and is calculated as follows:

$$Score(I(t_1)) = tf(t, t_1) \cdot idf(t_1, W_D) \tag{3}$$

where:
$tf(t, t_1)$ is the term frequency of 'query term t and its synonyms obtained from WordNet' in the article t_1, and
$idf(t_1, W_D)$ is the inverse document frequency of term t_1 in the whole of Wikipedia dump W_D. idf can be calculated as the following:

$$idf(t_1, W_D) = log \frac{N}{|\{d \in W_D : d \in t_1\}|} \tag{4}$$

where:
N is the total number of articles in Wikipedia dump, and
$|\{d \in W_D : d \in t_1\}|$ is the number of articles where the term t_1 appears.

The intuition behind the in-link score is to capture (1) the amount of similarity between the expansion term and the initial query term, and (2) the amount of useful information being provided by the expansion term with respect to QE, i.e, whether the expansion term is common or rare across whole Wikipedia dump.

Elaborating on the above two points, the term frequency tf provides the semantic similarity between the initial query term and the expansion term, whereas idf provides score for the rareness of an expansion term. The latter assigns lower priority to the stop words (common terms) in Wikipedia articles (e.g., Main page, contents, edit, References, Help, About Wikipedia, etc.). In Wikipedia both common terms and expansion terms are hyper-links of the query term article; the idf helps in removing these common hyper-links present in all the articles of the candidate expansion terms.

After assigning an in-link score to each expanded term, for each term in the initial query, we select top n terms based on their in-link scores. These top n terms form the intermediate expanded query. After this, these intermediate terms are re-weighted using correlation score (as described in Sec. 3.4). Top m terms chosen on the basis of correlation score become one part of the expanded query. The other part is obtained from WordNet as described next.

3.3 QE using WordNet

After preprocessing of the initial query, the individual terms and phrases obtained as keywords are searched in WordNet for QE. While extracting semantically similar terms from WordNet, more priority is given to the phrases in the query than the individual terms. Specifically, phrases (formed by two consecutive words) are looked up first in WordNet for expansion. Only when no entity is found in WordNet corresponding to a phrase, its individual terms are looked up separately in WordNet. It should be noted that phrases are considered only at the time of finding semantically similar terms from WordNet.

When querying for semantically similar terms from WordNet, only synonym and hyponyms sets of the query term are considered as candidate expansion terms. Here synonyms and hyponyms are fetched at two levels, i.e., for an initial query term Q_i, at level one its synonyms, denoted x_i, are considered, and, at level two, synonyms of x_is are considered as shown in Fig.5. The final synonym set used for QE is the union of level one and level two synonyms. Hyponyms are also fetched similarly at two levels.
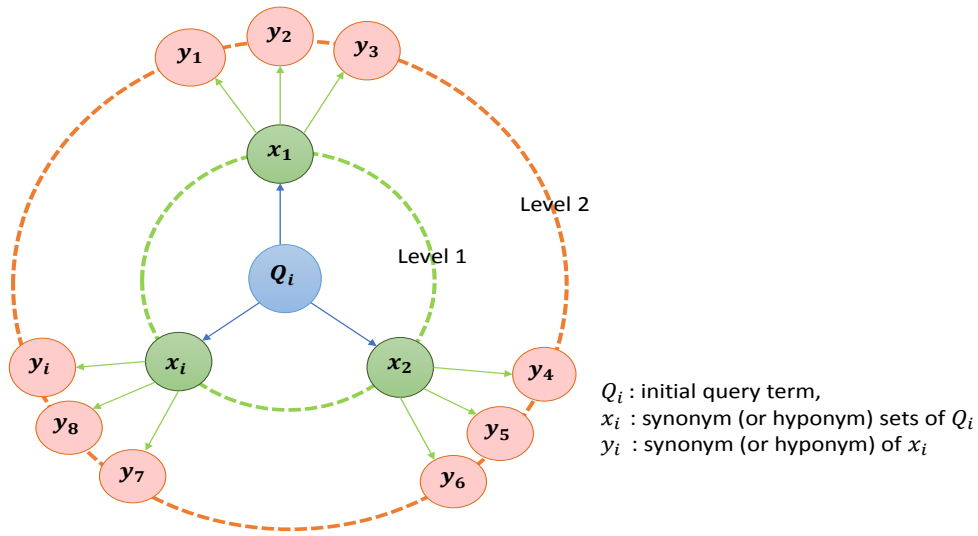

Fig. 5: Initial query term and its two level synonyms (or hyponyms) sets

After fetching synonyms and hyponyms at two levels, a wide range of semantically similar terms are obtained. Next, we rank these terms using $tf \cdot idf$:

$$Score(t_1) = tf(t_1, t) \cdot idf(t_1, W_D) \tag{5}$$

where:
t is the initial query term,
t_1 is an expanded term,
$tf(t_1, t)$ is the term frequency of expanded term t_1 in the Wikipedia article of query term t, and
$idf(t_1, W_D)$ is the inverse document frequency of term t_1 in whole Wikipedia dump W_D.
idf is calculated as given in Eq. 4.

After ranking expanded terms based on the above score, we collect the top n terms as the intermediate expanded query. These intermediate terms are re-weighted using correlation score. Top m terms chosen on the basis of correlation score (as described in Sec. 3.4) become the second part of the expanded query. The first part being obtained from Wikipedia as described before.

3.4 Re-weighting Expanded Terms

So far, a set of candidate expansion terms have been obtained, where each expansion term is strongly connected to an individual query term or phrase. These terms have been assigned weights

using in-link score (for terms obtained from Wikipedia) and $tf.idf$ score (for terms terms obtained from WordNet). However, this may not properly capture the relationship of the expansion term to the query as a whole. For example, the word "technology" is frequently associated with the word "information". Here, expanding the query term "technology" with "information" might work well for some queries such as "engineering technology", "science technology" and "educational technology" but might not work well for others such as "music technology", "food technology", and "financial technology". This problem has also been discussed in reference [13]. To resolve this language ambiguity problem, we re-weight expanded terms using correlation score [79, 103]. The logic behind doing so is that if an expansion feature is correlated to several individual query terms, then the chances are high that it will be correlated to the query as a whole as well.

The correlation score is described as follows. Let q be the original query and let t_1 be a candidate expansion term. The correlation score of t_1 with q is calculated as:

$$
\begin{aligned}
C_{q,t_1} &= \frac{1}{|q|} \sum_{t \in q} c_{t,t_1} \\
&= \frac{1}{|q|} \sum_{t \in q} w_{t,a_t} \cdot w_{t_1,a_t}
\end{aligned}
\tag{6}
$$

where:
c_{t,t_1} denotes correlation (similarity) score between terms t and t_1, and
w_{t,a_t} (w_{t_1,a_t}) is the weight of the term t (t_1) in the article a_t of term t.
The weight of the term t in its article $t(a_t)$, denoted w_{t,a_t} (w_{t_1,a_t} is similarly defined), is computed as:

$$
\begin{aligned}
w_{t,a_t} &= tf(t, a_t) \cdot itf(t, a_q) \\
&= tf(t, a_t) \cdot log \frac{T}{|T_{a_t}|}
\end{aligned}
\tag{7}
$$

where:
$tf(t, a_t)$ is the term frequency of term t in its article a_t,
a_q denotes all Wikipedia articles corresponding to the terms in the original query q,
$itf(t, a_q)$ is the inverse term frequency of term t associated with a_q,
T is the frequency of term t in all the Wikipedia articles in set a_q, and
$|T_{a_t}|$ is the frequency of term t in the article a_t.

After assigning the correlation score to expansion terms, we collect the top m terms from both data sources to form the final set of expanded terms.

4 Experimental Setup

In order to evaluate the proposed WWQE approach, the experiments were carried out on a large number of queries (50) from FIRE ad-hoc test collections [50]. As real life queries are short, we used only the title field of all queries. We used Brill's tagger to assign a POS tag to each query term for extracting the phrase and individual word. These phrase and individual words have been used for QE. We used the most recent Windows version of WordNet 2.1 to extract two level of synsets terms and Wikipedia for in-links extraction for QE.

We use the Wikipeia Dump (also known as 'WikiDump') for in-link extraction. Wikipedia dump contains every Wikipedia article in XML format. As an open source project, the Wikipedia dump can be download from *https://dumps.wikimedia.org/*. We download the English Wikipedia dump titled "enwiki-20170101-pages-articles-multistream.xml" of January 2017.

We compare the performance of our query expansion technique with several existing weighting models as described in Sec.4.2.

4.1 Dataset

We use the well known benchmark dataset Forum for Information Retrieval Evaluation (FIRE) [50] to evaluate our proposed WWQE approach. Table 1 summarizes the dataset used. FIRE collections

consists of a very large set of documents on which IR is done, a set of questions (called topics) and the right answers (called relevance judgments) stating relevance of documents to the corresponding topic(s). The FIRE dataset consists of a large collection of newswire articles from two sources namely BDnews24 [15] and The Telegraph [8] provided by Indian Statistical Institute Kolkata, India.

Table 1: Statistics of experimental corpora

Corpus	Source	Size	# of docs	Queries
FIRE	FIRE 2011 (English)	1.76 GB	3,92,577	126 - 175

4.2 Evaluation Metrics

We used the TERRIER[10] retrieval system for our all experimental evaluation. We use the title field of the topics in FIRE dataset. For indexing the documents, first stopwords are removed, then Porter's Stemmer is used for stemming the documents. All experimental evaluations are based on the unigram word assumption, i.e., all documents and queries in the corpus are indexed using single terms. We did not use any phrase or positional information. To compare the effectiveness of our expansion technique, we used the following weighting models: IFB2 a probabilistic divergence from randomness (DFR) model [6], BM25 model of Okapi [82], Laplace's law of succession I(n)L2 [90], Log-logistic DFR model LGD [27], Divergence from Independence model DPH [5] and Standard tf.idf model. The Parameters for these models were set to the default values in TERRIER.

We evaluate the results on standard evaluation metrics: MAP(mean average precision), GM_MAP (geometric mean average precision), P@10 (precision at top 10 ranks), P@20, P@30, bpref (binary preference) and the overall recall (number of relevant documents retrieved). Additionally, we report the percentage improvement in MAP over the baseline (non expanded query) for each expansion method.

5 Experimental Results

The aim of our experiments is to explore the effectiveness of the proposed Wikipedia-WordNet based QE technique (WWQE) by comparing it with the three baselines on popular weighting models and evaluation metrics. The comparison was done over three baselines: (i) unexpanded query, (ii) query expansion using Wikipedia alone, and (iii) query expansion using WordNet alone. Comparative analysis is shown in Tables 2, 3 and 4.

Table 4 shows performance comparison of the proposed WWQE technique over popular weighting models in the context of MAP, GM_MAP, P@10, P@20, P@30 and relevant return. The table shows that the proposed WWQE technique is compatible with the existing popular weighting models and it also improves the information retrieval effectively. It also shows the relative percentage improvements (within parentheses) of various standard evaluation metrics measured against no expansion. By using the proposed query expansion technique (WWQE), the weighting models improve the MAP up to 24% and GM_MAP by 48%. Based on the results presented in Table 4 we can say that in the context of all evaluation parameters, the proposed QE technique performs well with all weighting models.

Figure 6 shows the comparative analysis of precision-recall curve of WWQE technique with various weighting models. This graph plots the interpolated precision of an IR system using 11 standard cutoff values from the Recall levels, i.e {0, 0.1, 0.2, 0.3, ...,1.0}. These graphs are widely used to evaluate IR systems that return ranked documents (i.e., averaging and plotting retrieval results). Comparisons are best made in three different recall ranges: 0 to 0.2, 0.2 to 0.8, and 0.8 to 1. These ranges characterize high precision, middle recall, and high recall performance, respectively. Based on the graph presented in Figures 6a and 6b, we arrive at the conclusion that P-R curve of

---

10 http://terrier.org/

Table 2: Comparison of QE using Wikipedia alone on popular models with top 30 expansion terms on the FIRE Dataset

Model Performance Without Query Expansion						
Method	MAP	GM_MAP	P@10	P@20	P@30	#rel_ret
IFB2	0.2765	0.1907	0.3660	0.3560	0.3420	2330
I(n)L2	0.2979	0.2023	0.4280	0.3900	0.3553	2322
LGD	0.2909	0.1974	0.4100	0.3710	0.3420	2309
DPH	0.3133	0.2219	0.4540	0.4040	0.3653	2338
BM25	0.3163	0.2234	0.4600	0.3970	0.3660	2343
TF_IDF	0.3183	0.2261	0.4560	0.4010	0.3707	2340
Model Performance With QE using Wikipedia alone						
IFB2	0.3166	0.2498	0.4162	0.3969	0.3623	2420
	(↑**14.5%**)	(↑**30.99%**)	(↑**13.72%**)	(↑**11.49%**)	(↑5.94%)	(↑3.86%)
I(n)L2	0.3317	0.2628	0.4425	0.4012	0.3892	2432
	(↑11.35%)	(↑29.91%)	(↑3.39%)	(↑2.87%)	(↑**9.54%**)	(↑4.74%)
LGD	0.3248	0.2535	0.4432	0.3901	0.3639	2428
	(↑11.65%)	(↑28.42%)	(↑8.1%)	(↑5.15%)	(↑6.4%)	(↑**5.15%**)
DPH	0.3291	0.2598	0.4667	0.4127	0.3783	2423
	(↑5.04%)	(↑17.08%)	(↑2.8%)	(↑2.15%)	(↑3.56%)	(↑3.64%)
BM25	0.3304	0.2501	0.4723	0.4044	0.3717	2421
	(↑4.46%)	(↑11.95%)	(↑2.67%)	(↑1.86%)	(↑1.56%)	(↑3.33%)
TF_IDF	0.3315	0.2572	0.4691	0.4123	0.3875	2422
	(↑4.15%)	(↑13.75%)	(↑2.87%)	(↑2.82%)	(↑4.53%)	(↑3.5%)

Table 3: Comparison of QE using WordNet alone on popular models with top 30 expansion terms on the FIRE Dataset

Model Performance Without Query Expansion						
Method	MAP	GM_MAP	P@10	P@20	P@30	#rel_ret
IFB2	0.2765	0.1907	0.3660	0.3560	0.3420	2330
I(n)L2	0.2979	0.2023	0.4280	0.3900	0.3553	2322
LGD	0.2909	0.1974	0.4100	0.3710	0.3420	2309
DPH	0.3133	0.2219	0.4540	0.4040	0.3653	2338
BM25	0.3163	0.2234	0.4600	0.3970	0.3660	2343
TF_IDF	0.3183	0.2261	0.4560	0.4010	0.3707	2340
Model Performance With QE using WordNet alone						
IFB2	0.2901	0.2113	0.3817	0.3693	0.3521	2361
	(↑4.92%)	(↑10.8%)	(↑**4.29%**)	(↑3.74%)	(↑2.95%)	(↑1.33%)
I(n)L2	0.3112	0.2246	0.4373	0.3972	0.3648	2358
	(↑4.46%)	(↑**11.02%**)	(↑2.17%)	(↑1.85%)	(↑2.67%)	(↑**1.55%**)
LGD	0.3101	0.2177	0.4111	0.3872	0.3513	2327
	(↑**6.6%**)	(↑10.28%)	(↑0.27%)	(↑**4.37%**)	(↑2.72%)	(↑0.78%)
DPH	0.3178	0.2295	0.4627	0.4105	0.3712	2359
	(↑1.43%)	(↑3.42%)	(↑1.92%)	(↑1.61%)	(↑1.62%)	(↑0.89%)
BM25	0.3199	0.2301	0.4612	0.3999	0.3725	2353
	(↑1.14%)	(↑3%)	(↑0.26%)	(↑0.73%)	(↑1.78%)	(↑0.43%)
TF_IDF	0.3203	0.2312	0.4597	0.4098	0.3827	2345
	(↑0.63%)	(↑2.26%)	(↑0.84%)	(↑2.19%)	(↑**3.24%**)	(↑0.21%)

the various weighting models have nearly the same retrieval result with or without QE respectively. Therefor we can say that for improving the information retrieval in QE, choice of the weighting models is not so important. The importance lies in the choice of technique used for selecting the relevant expansion terms. The relevant expansion terms, in turn, come from data sources. Hence, the data sources also play an important role for effective QE. This conclusion also supports our proposed WWQE technique where we select the expansion terms on the basis of individual term weighting as well as assign a correlation score on the basis of entire query.

Figure 7 compares the performance of WWQE expansion technique with P-R curve using popular weighting models individually. Graphs in the figure show the improvement in retrieval results of WWQE technique when compared with the original unexpanded queries.

Figure 8 compares the WWQE technique in terms of precision, bpref and P@5 with various weighting models on FIRE dataset in comparison to unexpanded queries. Here, precision shows the ability of a system to present only relevant documents. P@5 measures the precision over top

Table 4: Comparison of proposed WWQE technique on popular models with top 30 expansion terms on the FIRE Dataset

Model Performance Without Query Expansion						
Method	MAP	GM_MAP	P@10	P@20	P@30	#rel_ret
IFB2	0.2765	0.1907	0.3660	0.3560	0.3420	2330
I(n)L2	0.2979	0.2023	0.4280	0.3900	0.3553	2322
LGD	0.2909	0.1974	0.4100	0.3710	0.3420	2309
DPH	0.3133	0.2219	0.4540	0.4040	0.3653	2338
BM25	0.3163	0.2234	0.4600	0.3970	0.3660	2343
TF_IDF	0.3183	0.2261	0.4560	0.4010	0.3707	2340
Model Performance With Proposed Query Expansion Technique						
IFB2	0.3439	0.2835	0.4660	0.4400	0.4040	2554
	(↑**24.38%**)	(↑**48.66%**)	(↑**27.49%**)	(↑**23.60%**)	(↑18.13%)	(↑9.61%)
I(n)L2	0.3552	0.2933	0.4900	0.4560	0.4200	2583
	(↑19.23%)	(↑44.98%)	(↑14.48%)	(↑16.92%)	(↑18.21%)	(↑**11.24%**)
LGD	0.3460	0.2855	0.4900	0.4460	0.4187	2566
	(↑18.94%)	(↑44.63%)	(↑19.51%)	(↑20.21%)	(↑**22.43%**)	(↑11.13%)
DPH	0.3497	0.2902	0.4940	0.4490	0.4113	2565
	(↑11.62%)	(↑30.78%)	(↑8.81%)	(↑11.14%)	(↑12.59%)	(↑9.71%)
BM25	0.3508	0.2878	0.5160	0.4490	0.4093	2560
	(↑10.91%)	(↑28.83%)	(↑12.17%)	(↑13.10%)	(↑11.83%)	(↑9.26%)
TF_IDF	0.3521	0.2896	0.5100	0.4520	0.4120	2561
	(↑10.62%)	(↑27.95%)	(↑11.84%)	(↑12.72%)	(↑11.14%)	(↑9.44%)

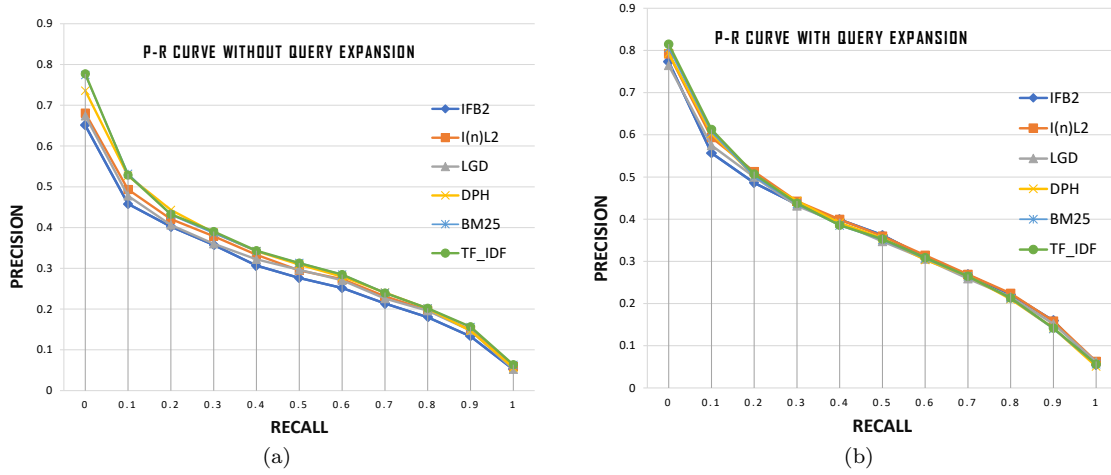(a)   (b)

Fig. 6: Comparative analysis of Precision-Recall curve of proposed QE technique with various weighting models on FIRE dataset.

5 documents retrieved and bpref measures a preference relation about how many judged relevant documents are ranked before judged irrelevant documents.

Figure 9 compare the WWQE technique in terms of MAP, bpref and P@5 with baseline (unexpanded), QE using WordNet alone and QE using Wikipedia alone. IFB2 model is used for term weighting for experimental evaluation.

After evaluating the performance of the proposed QE technique on several popular evaluation metrics, it can be concluded that the proposed QE technique (WWQE) performs well with all weighting models on several evaluation parameters. Therefor, the proposed WWQE technique is effective in improving information retrieval results.

6 Conclusion

This article presents a novel Wikipedia WordNet based Query Expansion (WWQE) technique that considers the individual terms and phrases as the expansion terms. Proposed technique employs a two level strategy to select terms from WordNet. First, it fetches synsets of the initial query terms.
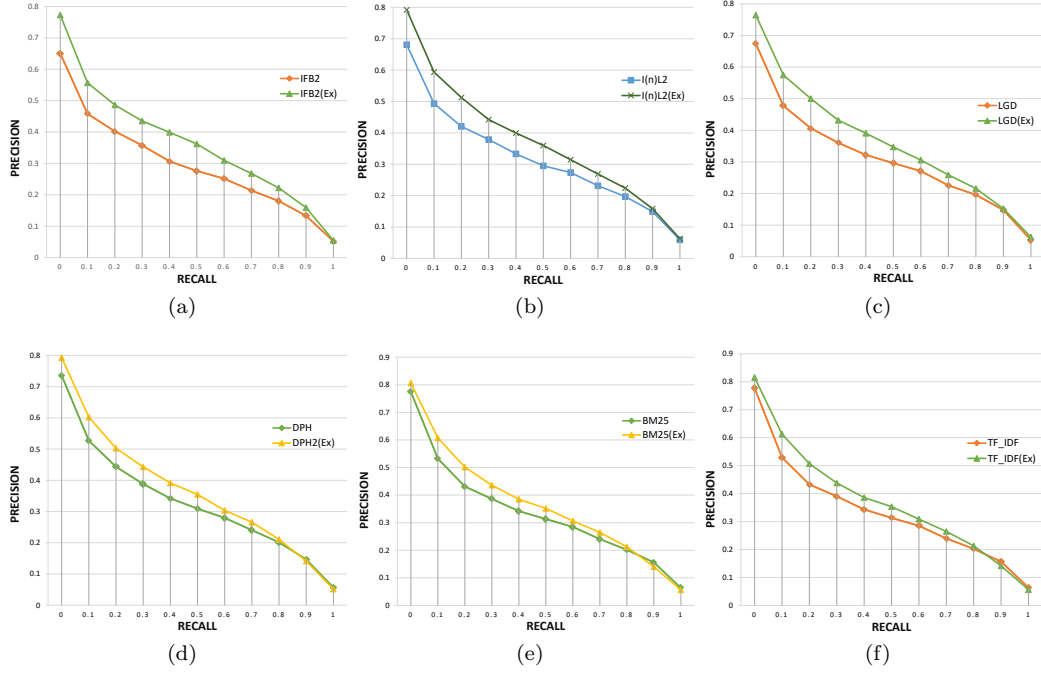
Fig. 7: Comparative analysis of Precision-Recall curve of WWQE technique with popular weighting models individually on FIRE dataset.
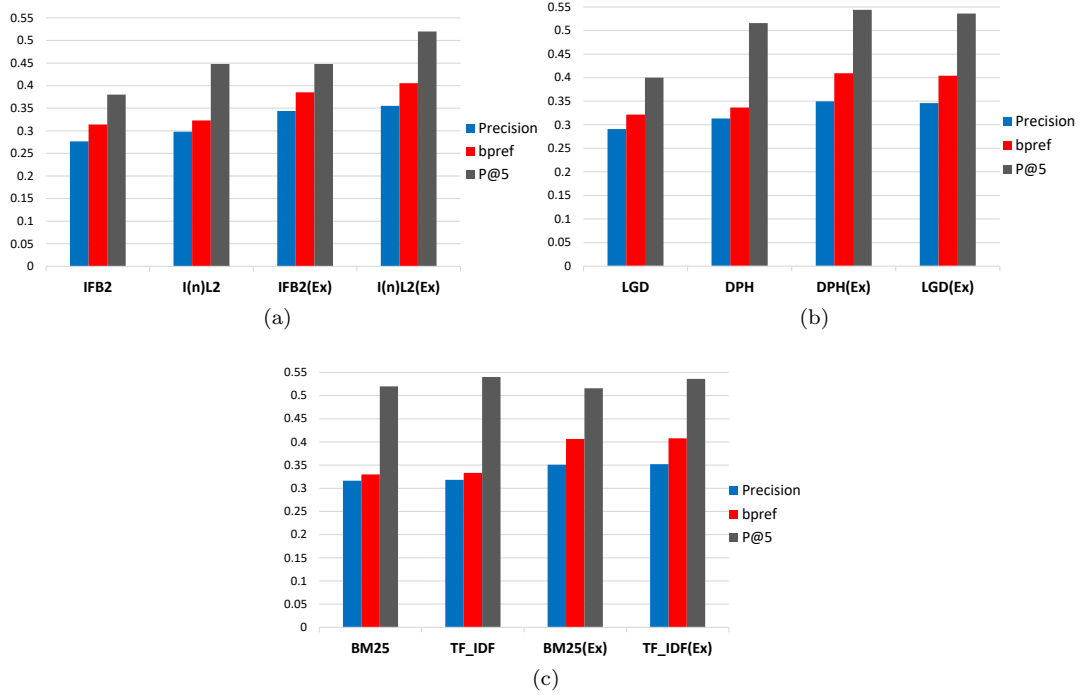

Fig. 8: Comparative analysis of WWQE technique in terms of precision, bpref and P@5 with various weighting models on FIRE dataset.

Then, it extracts sysnets of these synsets. In order to score the expansion term on Wikipedia, we proposed a new weighting score named as in-link score. The in-link score assigns a score to each expansion term extracted from Wikipedia, and tf-idf based scoring system is used to assign a score to expansion terms extracted from WordNet. After assigning score to individual query terms, we
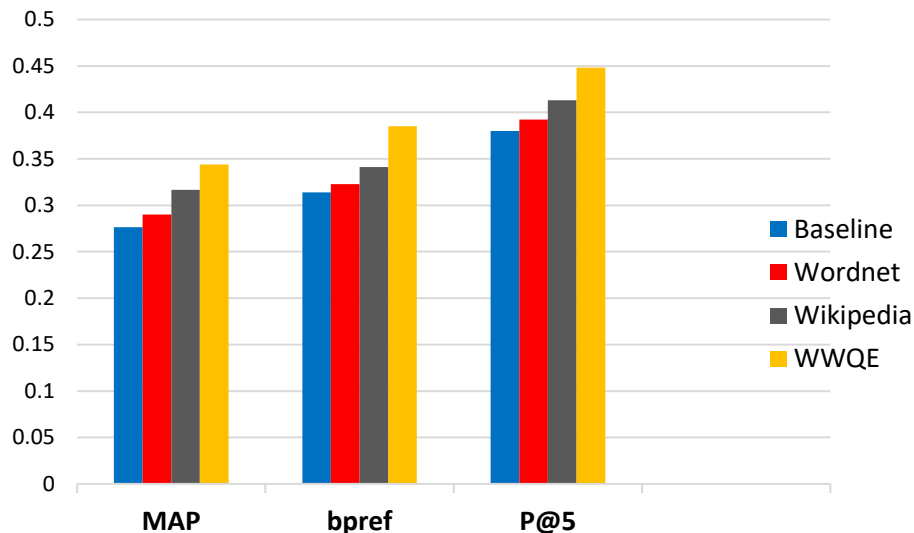
Fig. 9: Comparative analysis of WWQE technique with baseline, WordNet and Wikipedia

further re-weight the selected expansion terms using correlation score with respect to the entire query. The combination of the two data sources works well for extracting relevant expansion terms and the proposed QE technique performs well with these terms on several weighting models. It also yields better results when compared to the the two methods individually. The results on the basis of several evaluation metrics on FIRE dataset demonstrate the effectiveness of our proposed QE technique in the field of information retrieval. The proposed query expansion technique improves the IR effectively on evaluation with several popular weighting models.

References

1. Aggarwal, N., Buitelaar, P.: Query expansion using wikipedia and dbpedia. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
2. Agichtein, E., Lawrence, S., Gravano, L.: Learning to find answers to questions on the web. ACM Transactions on Internet Technology (TOIT) **4**(2), 129–162 (2004)
3. Al-Shboul, B., Myaeng, S.H.: Wikipedia-based query phrase expansion in patent class search. Information retrieval **17**(5-6), 430–451 (2014)
4. ALMasri, M., Berrut, C., Chevallet, J.P.: Wikipedia-based semantic query enrichment. In: Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval, pp. 5–8. ACM (2013)
5. Amati, G., Amodeo, G., Bianchi, M., Gaibisso, C., Gambosi, G.: Fub, iasi-cnr and university of tor vergata at trec 2008 blog track. Tech. rep., FONDAZIONE UGO BORDONI ROME (ITALY) (2008)
6. Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS) **20**(4), 357–389 (2002)
7. Arguello, J., Elsas, J.L., Callan, J., Carbonell, J.G.: Document representation and query expansion models for blog recommendation. ICWSM **2008**(0), 1 (2008)
8. Arup Sarkar: Daily newspaper. https://www.telegraphindia.com/ (2018). [Online; accessed 29-August-2018]
9. Augenstein, I., Gentile, A.L., Norton, B., Zhang, Z., Ciravegna, F.: Mapping keywords to linked data resources for automatic query expansion. In: Extended Semantic Web Conference, pp. 101–112. Springer (2013)
10. Azad, H.K., Deepak, A.: Query expansion techniques for information retrieval: a survey. arXiv preprint arXiv:1708.00247 (2017)
11. Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query recommendation using query logs in search engines. In: International Conference on Extending Database Technology, pp. 588–596. Springer (2004)
12. Baeza-Yates, R., Tiberi, A.: Extracting semantic relations from query logs. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 76–85. ACM (2007)
13. Bai, J., Nie, J.Y., Cao, G., Bouchard, H.: Using query contexts in information retrieval. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 15–22. ACM (2007)
14. Bai, J., Song, D., Bruza, P., Nie, J.Y., Cao, G.: Query expansion using term relationships in language models for information retrieval. In: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 688–695. ACM (2005)
15. Bangladesh News 24 Hours Ltd.: Online newspaper. https://bdnews24.com/ (2018). [Online; accessed 29-August-2018]
16. Bhatia, S., Majumdar, D., Mitra, P.: Query suggestions in the absence of query logs. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 795–804. ACM (2011)

17. Bhogal, J., Macfarlane, A., Smith, P.: A review of ontology based query expansion. Information processing & management **43**(4), 866–886 (2007)
18. Biancalana, C., Gasparetti, F., Micarelli, A., Sansonetti, G.: Social semantic query expansion. ACM Transactions on Intelligent Systems and Technology (TIST) **4**(4), 60 (2013)
19. Billerbeck, B., Scholer, F., Williams, H.E., Zobel, J.: Query expansion using associated queries. In: Proceedings of the twelfth international conference on Information and knowledge management, pp. 2–9. ACM (2003)
20. Bilotti, M.W., Katz, B., Lin, J.: What works better for question answering: Stemming or morphological query expansion. In: Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR, vol. 2004, pp. 1–3 (2004)
21. Bouadjenek, M.R., Hacid, H., Bouzeghoub, M., Vakali, A.: Persador: Personalized social document representation for improving web search. Information Sciences **369**, 614–633 (2016)
22. Brill, E.: Penn treebank tagger. Copyright by MIT and University of Pennsylvania
23. Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., Li, H.: Context-aware query suggestion by mining click-through and session data. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 875–883. ACM (2008)
24. Carpineto, C., De Mori, R., Romano, G., Bigi, B.: An information-theoretic approach to automatic query expansion. ACM Transactions on Information Systems (TOIS) **19**(1), 1–27 (2001)
25. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Computing Surveys (CSUR) **44**(1), 1 (2012)
26. Chakraborty, G., Pagolu, M.K.: Analysis of unstructured data: Applications of text analytics and sentiment mining. In: SAS global forum, pp. 1288–2014 (2014)
27. Clinchant, S., Gaussier, E.: Information-based models for ad hoc ir. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 234–241. ACM (2010)
28. Collins-Thompson, K., Callan, J.: Query expansion using random walk models. In: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 704–711. ACM (2005)
29. Crouch, C.J., Yang, B.: Experiments in automatic statistical thesaurus construction. In: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 77–88. ACM (1992)
30. Cui, H., Wen, J.R., Nie, J.Y., Ma, W.Y.: Probabilistic query expansion using query logs. In: Proceedings of the 11th international conference on World Wide Web, pp. 325–332. ACM (2002)
31. Cui, H., Wen, J.R., Nie, J.Y., Ma, W.Y.: Query expansion by mining user logs. IEEE Transactions on knowledge and data engineering **15**(4), 829–839 (2003)
32. Dalton, J., Dietz, L., Allan, J.: Entity query feature expansion using knowledge base links. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pp. 365–374. ACM (2014)
33. Dang, V., Croft, B.W.: Query reformulation using anchor text. In: Proceedings of the third ACM international conference on Web search and data mining, pp. 41–50. ACM (2010)
34. Elsas, J.L., Arguello, J., Callan, J., Carbonell, J.G.: Retrieval and feedback models for blog feed search. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 347–354. ACM (2008)
35. Fang, H.: A re-examination of query expansion using lexical resources. proceedings of ACL-08: HLT pp. 139–147 (2008)
36. Fitzpatrick, L., Dent, M.: Automatic feedback using past queries: social searching? In: ACM SIGIR Forum, vol. 31, pp. 306–313. ACM (1997)
37. Fonseca, B.M., Golgher, P., Pôssas, B., Ribeiro-Neto, B., Ziviani, N.: Concept-based interactive query expansion. In: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 696–703. ACM (2005)
38. Gan, L., Hong, H.: Improving query expansion for information retrieval using wikipedia. International Journal of Database Theory and Application **8**(3), 27–40 (2015)
39. Gauch, S., Wang, J., Rachakonda, S.M.: A corpus analysis approach for automatic query expansion and its extension to multiple databases. ACM Transactions on Information Systems (TOIS) **17**(3), 250–269 (1999)
40. Ghorab, M.R., Zhou, D., OConnor, A., Wade, V.: Personalised information retrieval: survey and classification. User Modeling and User-Adapted Interaction **23**(4), 381–443 (2013)
41. Gong, Z., Cheang, C.W., et al.: Multi-term web query expansion using wordnet. In: International Conference on Database and Expert Systems Applications, pp. 379–388. Springer (2006)
42. Gonzalo, J., Verdejo, F., Chugur, I., Cigarran, J.: Indexing with wordnet synsets can improve text retrieval. arXiv preprint cmp-lg/9808002 (1998)
43. Guisado-Gámez, J., Prat-Pérez, A., Larriba-Pey, J.L.: Query expansion via structural motifs in wikipedia graph. arXiv preprint arXiv:1602.07217 (2016)
44. He, B., Ounis, I.: Combining fields for query expansion and adaptive query expansion. Information processing & management **43**(5), 1294–1307 (2007)
45. Hu, J., Deng, W., Guo, J.: Improving retrieval performance by global analysis. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, vol. 2, pp. 703–706. IEEE (2006)
46. Hua, X.S., Yang, L., Wang, J., Wang, J., Ye, M., Wang, K., Rui, Y., Li, J.: Clickage: towards bridging semantic and intent gaps via mining click logs of search engines. In: Proceedings of the 21st ACM international conference on Multimedia, pp. 243–252. ACM (2013)
47. Huang, C.K., Chien, L.F., Oyang, Y.J.: Relevant term suggestion in interactive web search based on contextual information in query session logs. Journal of the Association for Information Science and Technology **54**(7), 638–649 (2003)
48. Huang, J., Efthimiadis, E.N.: Analyzing and evaluating query reformulation strategies in web search logs. In: Proceedings of the 18th ACM conference on Information and knowledge management, pp. 77–86. ACM (2009)
49. Hull, D.A., et al.: Stemming algorithms: A case study for detailed evaluation. JASIS **47**(1), 70–84 (1996)

50. Information Retrieval Society of India: Forum for information retrieval evaluation. http://fire.irsi.res.in/fire/static/data (2018). [Online; accessed 25-August-2018]
51. Jardine, N., van Rijsbergen, C.J.: The use of hierarchic clustering in information retrieval. Information storage and retrieval **7**(5), 217–240 (1971)
52. Jones, K.S.: Automatic keyword classification for information retrieval (1971)
53. Karan, M., Šnajder, J.: Evaluation of manual query expansion rules on a domain specific faq collection. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 248–253. Springer (2015)
54. Kraft, R., Zien, J.: Mining anchor text for query refinement. In: Proceedings of the 13th international conference on World Wide Web, pp. 666–674. ACM (2004)
55. Krovetz, R.: Viewing morphology as an inference process. In: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 191–202. ACM (1993)
56. Kuzi, S., Shtok, A., Kurland, O.: Query expansion using word embeddings. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 1929–1932. ACM (2016)
57. Lee, K.S., Croft, W.B., Allan, J.: A cluster-based resampling method for pseudo-relevance feedback. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 235–242. ACM (2008)
58. Lemos, O.A., de Paula, A.C., Zanichelli, F.C., Lopes, C.V.: Thesaurus-based automatic query expansion for interface-driven code search. In: Proceedings of the 11th Working Conference on Mining Software Repositories, pp. 212–221. ACM (2014)
59. Li, Y., Luk, W.P.R., Ho, K.S.E., Chung, F.L.K.: Improving weak ad-hoc queries using wikipedia asexternal corpus. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 797–798. ACM (2007)
60. Liu, S., Liu, F., Yu, C., Meng, W.: An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 266–272. ACM (2004)
61. Liu, Y., Li, C., Zhang, P., Xiong, Z.: A query expansion algorithm based on phrases semantic similarity. In: Information Processing (ISIP), 2008 International Symposiums on, pp. 31–35. IEEE (2008)
62. Lu, M., Sun, X., Wang, S., Lo, D., Duan, Y.: Query expansion via wordnet for effective code search. In: Software Analysis, Evolution and Reengineering (SANER), 2015 IEEE 22nd International Conference on, pp. 545–549. IEEE (2015)
63. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
64. Maron, M.E., Kuhns, J.L.: On relevance, probabilistic indexing and information retrieval. Journal of the ACM (JACM) **7**(3), 216–244 (1960)
65. McBryan, O.A.: Genvl and wwww: Tools for taming the web. In: Proceedings of the first international world wide web conference, vol. 341. Geneva (1994)
66. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to wordnet: An on-line lexical database. International journal of lexicography **3**(4), 235–244 (1990)
67. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM conference on Information and knowledge management, pp. 509–518. ACM (2008)
68. Minker, J., Wilson, G.A., Zimmerman, B.H.: An evaluation of query expansion by the addition of clustered terms for a document retrieval system. Information Storage and Retrieval **8**(6), 329–348 (1972)
69. Moreau, F., Claveau, V., Sébillot, P.: Automatic morphological query expansion using analogy-based machine learning. In: European Conference on Information Retrieval, pp. 222–233. Springer (2007)
70. Natsev, A.P., Haubold, A., Tešić, J., Xie, L., Yan, R.: Semantic concept-based query expansion and re-ranking for multimedia retrieval. In: Proceedings of the 15th ACM international conference on Multimedia, pp. 991–1000. ACM (2007)
71. Navigli, R.: Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) **41**(2), 10 (2009)
72. Navigli, R., Velardi, P.: Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. IEEE transactions on pattern analysis and machine intelligence **27**(7), 1075–1086 (2005)
73. Paice, C.D.: An evaluation method for stemming algorithms. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 42–50. Springer-Verlag New York, Inc. (1994)
74. Pal, A.R., Saha, D.: Word sense disambiguation: A survey. arXiv preprint arXiv:1508.01346 (2015)
75. Pal, D., Mitra, M., Bhattacharya, S.: Exploring query categorisation for query expansion: A study. arXiv preprint arXiv:1509.05567 (2015)
76. Pal, D., Mitra, M., Datta, K.: Improving query expansion using wordnet. Journal of the Association for Information Science and Technology **65**(12), 2469–2478 (2014)
77. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)
78. Porter, M.F.: Implementing a probabilistic information retrieval system. Information Technology: Research and Development **1**(2), 131–156 (1982)
79. Qiu, Y., Frei, H.P.: Concept based query expansion. In: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 160–169. ACM (1993)
80. Riezler, S., Vasserman, A., Tsochantaridis, I., Mittal, V., Liu, Y.: Statistical machine translation for query expansion in answer retrieval. In: Annual Meeting-Association For Computational Linguistics, vol. 45, p. 464 (2007)
81. van Rijsbergen, C.J.: A theoretical basis for the use of co-occurrence data in information retrieval. Journal of documentation **33**(2), 106–119 (1977)
82. Robertson, S.E., Walker, S., Beaulieu, M., Gatford, M., Payne, A.: Okapi at trec-4. Nist Special Publication Sp pp. 73–96 (1996)
83. Rocchio, J.J.: Relevance feedback in information retrieval (1971)

84. Salton, G.: Automatic text processing: The transformation, analysis, and retrieval of. Reading: Addison-Wesley (1989)
85. Salton, G.: Developments in automatic text retrieval. science **253**(5023), 974–980 (1991)
86. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management **24**(5), 513–523 (1988)
87. Smeaton, A.F., Kelledy, F., O'Donnell, R.: Trec-4 experiments at dublin city university: Thresholding posting lists, query expansion with wordnet and pos tagging of spanish. Harman [6] pp. 373–389 (1995)
88. Soricut, R., Brill, E.: Automatic question answering using the web: Beyond the factoid. Information Retrieval **9**(2), 191–206 (2006)
89. Sun, R., Ong, C.H., Chua, T.S.: 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, pp. 382–389. ACM (2006)
90. Theodorescu, R.: Good, i. j.: The estimation of probabilities. an essay on modern bayesian methods. research monograph no. 30, the m. i. t. press cambridge 1965. 109 s., 7 tab., 115 lit. Biometrische Zeitschrift **10**(1), 87–87. DOI 10.1002/bimj.19680100118. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.19680100118
91. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pp. 173–180. Association for Computational Linguistics (2003)
92. Unger, C., Ngomo, A.C.N., Cabrio, E.: 6th open challenge on question answering over linked data (qald-6). In: Semantic Web Evaluation Challenge, pp. 171–177. Springer (2016)
93. Van Rijsbergen, C.J.: A non-classical logic for information retrieval. The computer journal **29**(6), 481–485 (1986)
94. Voorhees, E.M.: On expanding query vectors with lexically related words. In: TREC, pp. 223–232 (1993)
95. Voorhees, E.M.: Query expansion using lexical-semantic relations. In: SIGIR94, pp. 61–69. Springer (1994)
96. Wang, X., Zhai, C.: Learn from web search logs to organize search results. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 87–94. ACM (2007)
97. Wang, X., Zhai, C.: Mining term association patterns from search logs for effective query reformulation. In: Proceedings of the 17th ACM conference on Information and knowledge management, pp. 479–488. ACM (2008)
98. White, R.W., Ruthven, I., Jose, J.M.: A study of factors affecting the utility of implicit relevance feedback. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 35–42. ACM (2005)
99. Wikipedia contributors: Wikipedia — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Wikipedia (2018). [Online; accessed 7-May-2018]
100. Wu, H., Wu, W., Zhou, M., Chen, E., Duan, L., Shum, H.Y.: Improving search relevance for short queries in community question answering. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, WSDM '14, pp. 43–52. ACM, New York, NY, USA (2014). DOI 10.1145/2556195.2556239. URL http://doi.acm.org/10.1145/2556195.2556239
101. Wu, J., Ilyas, I., Weddell, G.: A study of ontology-based query expansion. Technical report CS-2011–04 (2011)
102. Xiong, C., Callan, J.: Query expansion with freebase. In: Proceedings of the 2015 International Conference on The Theory of Information Retrieval, pp. 111–120. ACM (2015)
103. Xu, J., Croft, W.B.: 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, pp. 4–11. ACM (1996)
104. Xu, Y., Jones, G.J., Wang, B.: Query dependent pseudo-relevance feedback based on wikipedia. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 59–66. ACM (2009)
105. Xue, G.R., Zeng, H.J., Chen, Z., Yu, Y., Ma, W.Y., Xi, W., Fan, W.: Optimizing web search using web clickthrough data. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management, pp. 118–126. ACM (2004)
106. Yin, Z., Shokouhi, M., Craswell, N.: Query expansion using external evidence. In: European Conference on Information Retrieval, pp. 362–374. Springer (2009)
107. Yu, C.T., Buckley, C., Lam, K., Salton, G.: A generalized term dependence model in information retrieval. Tech. rep., Cornell University (1983)
108. Zhang, J., Deng, B., Li, X.: Concept based query expansion using wordnet. In: Proceedings of the 2009 international e-conference on advanced science and technology, pp. 52–55. IEEE Computer Society (2009)
109. Zhang, Y., Clark, S.: Syntactic processing using the generalized perceptron and beam search. Computational linguistics **37**(1), 105–151 (2011)
110. Zhang, Z., Wang, Q., Si, L., Gao, J.: Learning for efficient supervised query expansion via two-stage feature selection. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 265–274. ACM (2016)