

# A New Approach for Query Expansion using Wikipedia and WordNet

Hiteshwar Kumar Azad\*, Akshay Deepak

*Dept. of Computer Science & Engineering  
National Institute of Technology Patna, India*

---

## Abstract

Query expansion (QE) is a well-known technique used to enhance the effectiveness of information retrieval. QE reformulates the initial query by adding similar terms that help in retrieving more relevant results. Several approaches have been proposed in literature producing quite favorable results, but they are not evenly favorable for all types of queries (individual and phrase queries). One of the main reasons for this is the use of the same kind of data sources and weighting scheme while expanding both the individual and the phrase query terms. As a result, the holistic relationship among the query terms is not well captured or scored. To address this issue, we have presented a new approach for QE using Wikipedia and WordNet as data sources. Specifically, Wikipedia gives rich expansion terms for phrase terms, while WordNet does the same for individual terms. We have also proposed novel weighting schemes for expansion terms: *in-link score* (for terms extracted from Wikipedia) and a tf-idf 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 concerning the entire query using correlation score. The proposed approach gains improvements of 24% on the MAP score and 48% on the GMAP score over unexpanded queries on the FIRE dataset. Experimental results achieve a significant improvement over individual expansion and other related state-of-the-art approaches. We also analyzed the effect on retrieval effectiveness of the proposed technique by varying the number of expansion terms.

*Keywords:* Query Expansion, Information Retrieval, WordNet, Wikipedia

---

## 1. Introduction

The Web is the largest information source available on the planet and it's growing day by day [32]. According to a recent survey [11] in Computerworld magazine, approximately 70–80 percent of all data available to enterprises/organizations is unstructured information, i.e., information that either is not organized in a pre-defined manner or does not have a pre-defined data model. This

---

\*Corresponding Author

Email addresses: [hiteshwar.cse15@nitp.ac.in](mailto:hiteshwar.cse15@nitp.ac.in) (Hiteshwar Kumar Azad ), [akshayd@nitp.ac.in](mailto:akshayd@nitp.ac.in) (Akshay Deepak)

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 set of relevant documents  $D$  to use different vocabulary and language styles while referring to the same concept. For example, terms 'buy' and 'purchase' have the same meaning; however, only one of these can be present in the documents' index while the other one can be the user's query term. This makes it difficult to retrieve the information actually wanted by the user [21, 28]. An effective strategy to fill this gap is to use the Query Expansion (QE) technique, which 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 [26]. In this sense, the approach for selection of expansion terms is equally important in comparison to what we do further with the expanded terms in order to retrieve the desired information. QE has a long research history in information retrieval (IR) [31, 39]. It has the potential to enhance IR's 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 [7], feedback documents (few top ranked documents are retrieved in response to the initial query) [25], or external knowledge resources [15].

References [10, 5] provide comprehensive surveys on data sources used for QE. Broadly, such sources can be classified into four categories: (i) documents used in the retrieval process [7] (e.g., corpus), (ii) hand-built knowledge resources [35] (e.g., WordNet<sup>1</sup>, ConceptNet<sup>2</sup>, thesaurus, ontologies), (iii) external text collections and resources [15] (e.g., Web, Wikipedia), and (iv) hybrid data sources [14].

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., [7, 16]). 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".

Hand-built knowledge-resources-based QE extracts knowledge from textual hand-built data sources such as dictionaries, thesauruses, ontologies, and the LOD cloud (e.g., [44, 47, 35]). Thesaurus-based QE can be either automatic or hand-built. One of the most famous hand-built thesauruses is WordNet. While it significantly improves the retrieval effectiveness of badly constructed queries, it does not show a lot of 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 nouns, 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 [15]. In such cases, QE approaches show

---

<sup>1</sup><https://wordnet.princeton.edu/>

<sup>2</sup><http://conceptnet5.media.mit.edu/>

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 [13] uses WordNet, an external corpus, and the top retrieved documents as data sources for QE. Some of the other research works for query processing based on hybrid resources are [14, 24, 18].

Among the above data sources, Wikipedia and WordNet are popular choices for semantic enrichment of the initial query [44, 35, 2]. They are also two of the most widely used knowledge resources in natural language processing. Wikipedia is the largest encyclopedia describing entities. 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 [44, 27, 35, 2] 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 tf-idf-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 and other related state-of-the-art approaches. We also analyzed the effect on retrieval effectiveness of the proposed technique by varying the number of expansion terms. The experiments were carried out on the FIRE<sup>3</sup> dataset using popular weighting models and evaluation metrics.

### 1.1. Contributions

The contributions of this paper are as follows:

- *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.* The proposed WWQE technique employs a two-level strategy to select terms from WordNet based on a proposed tf-idf-based weighting scheme. For fetching expansion

---

<sup>3</sup><http://fire.irsi.res.in/fire/static/data>

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 deal with the user’s initial query, the proposed WWQE technique selects both phrases and individual words. A phrase usually offers a richer context and has less ambiguity. Hence, the expansion terms retrieved in response to phrases are more effective than expansion terms in response to the non-phrase terms from the initial query.
- *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) and other state-of-the-art approaches. We also analyzed the effect on retrieval effectiveness of the proposed technique by varying the number of expansion terms.

### 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 the study in Section 6.

## 2. Related Work

There is a wealth of literature on query expansion in the area of information retrieval (IR). In the 1960s, Moron and Kuhns [31] were the first researchers who applied QE for literature indexing and searching in a mechanized library system. In 1971, Rocchio [39] put QE under the spotlight through the “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, cluster-based information retrieval, comparative analysis of term distribution, and automatic text processing [43, 22].

Recently, QE has been focused on because a lot of researchers are using QE techniques for working on personalized social bookmarking services [48], Question Answering over Linked Data

(QALD)<sup>4</sup>, Text Retrieval Conference (TREC)<sup>5</sup>, and Forum for Information Retrieval Evaluation (FIRE)<sup>6</sup> collections. They are also used heavily in web, desktop and email searches [34, 37, 33]. Many platforms provide a QE facility to end users, which can be turned on or off, e.g., WordNet<sup>7</sup>, ConceptNet<sup>8</sup>, Lucene<sup>9</sup>, Google Enterprise<sup>10</sup> and MySQL. Some surveys have previously been done on QE techniques. In 2007, Bhogal et al. [8] reviewed QE techniques using ontologies, which are domain-specific. Such techniques have also been described in book [30]. Carpineto and Romano [10] 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.

### 2.1. Use of WordNet as a Data Source for QE

WordNet is one of the most popular hand-built thesauruses and 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 can synsets of these synsets also have meanings similar to the query term and hence should they also be considered as potential expansion terms?
- When considering a synset of a query term, should only synonyms be considered or should other relations (i.e., hypernyms, hyponyms, holonyms, meronyms etc.) 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. Reference [44] added manually selected WordNet synsets for QE, but unfortunately no significant improvements were obtained. Reference [42] used synonyms of the initial query and assigns half the weight. Reference [27] used word sense to add synonyms, hyponyms and terms' WordNet glosses to expand the query. Their experiments yielded significant improvements on TREC datasets. Reference [47] used sense disambiguation of query terms to add synonyms for QE. During experimental evaluation, in response to the user's initial query, reference [47]'s method produced an improvement of around 7% in P@10 value over the CACM collection. Reference [19] used a set of Candidate Expansion Terms (CET)

---

<sup>4</sup><http://qald.sebastianwalter.org/>

<sup>5</sup><http://trec.nist.gov/>

<sup>6</sup><http://fire.irs1.res.in/>

<sup>7</sup><https://wordnet.princeton.edu/>

<sup>8</sup><http://conceptnet5.media.mit.edu/>

<sup>9</sup><http://lucene.apache.org/>

<sup>10</sup><https://enterprise.google.com/search/>

that included 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 [35] 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 [23] 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 [29] used WordNet for effective code search, where it was used to generate synonyms. These synonyms were used as query expansion terms. During experimental evaluation, their approach showed improvements in precision and recall values of 5% and 8% respectively.

In almost all the aforementioned studies, CETs were 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 assigned weights to the synonyms level-wise.

## 2.2. Use of Wikipedia as Data Source for QE

Wikipedia is freely available and is 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 has been 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., [25, 17, 2]). Reference [25] 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 [17] used link-based QE on Wikipedia and focused on anchor text. It also proposed a phrase-scoring function. Reference [46] 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 covers a narrow topic), and (3) broader queries (queries having neither an ambiguous nor specific meaning). They consolidated the expansion terms into the original query and evaluated these techniques using language modelling IR. Reference [2] uses Wikipedia for semantic enrichment of short queries based on in-link and out-link articles. Reference [14] proposes the 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 three 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. Here, terms are scored by a newly proposed tf-idf-based weighting scheme.
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.

### 3. Our Approach

We present a new way of QE combining Wikipedia and WordNet as data sources. In order to expand the user's query, first, we pre-process the initial query to extract the phrases and individual terms. These phrases and individual terms are used to obtain similar terms from Wikipedia and WordNet. After this, we assign the similarity score to these terms using the proposed weighting score. This is done separately for terms from Wikipedia and WordNet. Then, the top  $n$  terms from Wikipedia and WordNet are selected as expansion terms. These expansion terms are re-weighted based on the correlation score. Finally, we select the top  $m$  terms as the final set of expansion terms.

The proposed approach consists of four main steps: (i) Pre-processing of initial query, (ii) QE using Wikipedia, (iii) QE using WordNet, and (iv) Re-weighting of expanded terms. Figure 1 summarizes these steps.

#### 3.1. Pre-processing of Initial Query

In the Preprocessing step, Brill's tagger is used to lemmatize each query and assign a Part Of Speech (POS) to each word in the query. The POS tagging is done on the 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 [1]. Phrases usually offer richer context and have less ambiguity. Hence, the documents retrieved in response to the phrases from the initial query have more importance than the documents retrieved in response to the 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. Therefore, we give more priority to phrases in the query than the individual words when finding expansion terms from Wikipedia and WordNet.

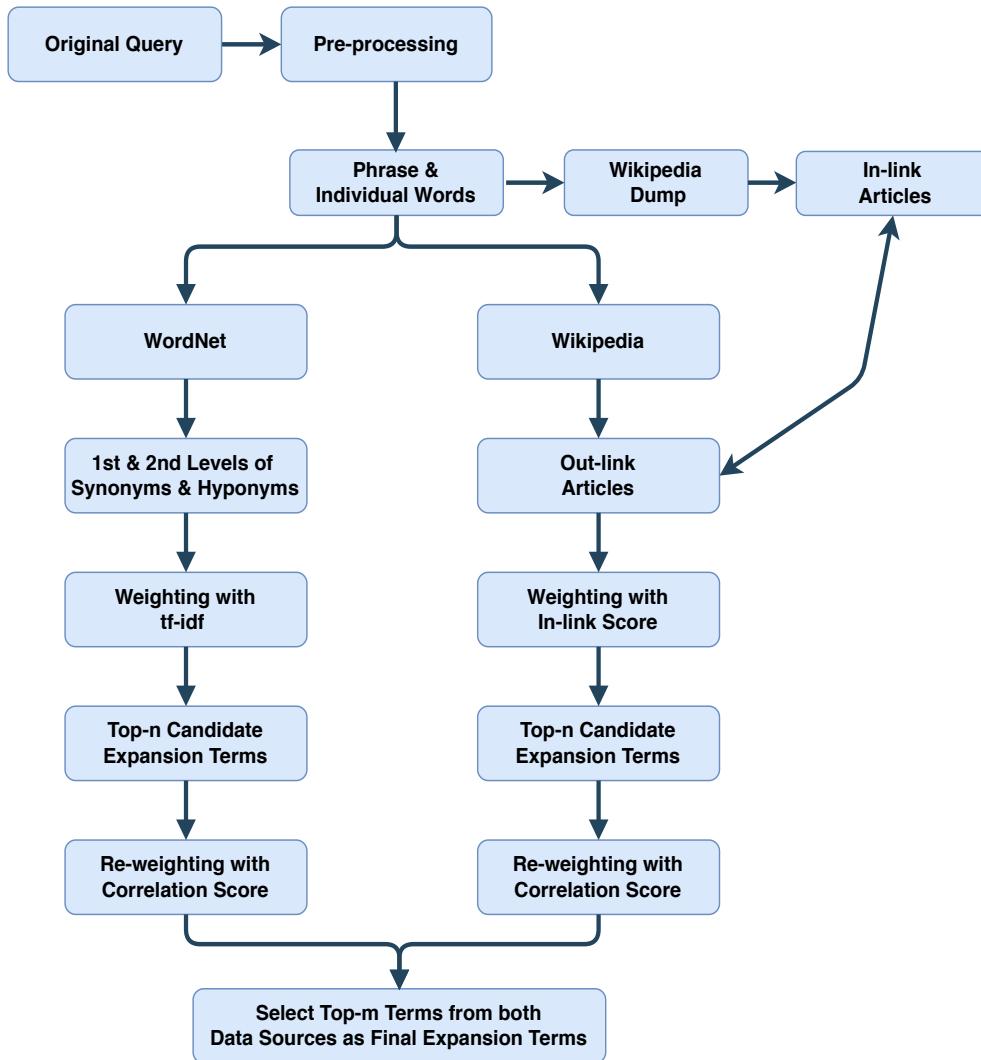


Figure 1: Steps involved in the proposed approach

For example, consider the following query (Query ID- 126) from the 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 vaccines
use on humans and animals , arrangements that are in place to
prevent scarcity / unavailability of the vaccine , and the
vaccines role in saving lives.</narr>
</top>

```

Multiple such queries in the standard SGML format are present in the query file of the FIRE

dataset. For extracting the root query, we extract the title from each query and tag it using the Stanford POS tagger library. For example, after POS tagging, the title of the above query looks like:

*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 nouns, adjectives, or verbs, or words with the cardinal number 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 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 a directed graph  $G(A, L)$ , where  $A$  and  $L$  indicate articles and links respectively. Each article  $x \in A$  effectively summarizes its entity, i.e., the title of  $x$ :  $\text{title}(x)$ , and provides links to the user to browse other related articles. In our work, we consider two types of links: in-links and out-links.

***In-links*** of  $x$ , denoted  $I(x)$ , is the set of articles that point to the article  $x$ . It can be defined as

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

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*** of  $x$ , denoted  $O(x)$  is the set of articles that  $x$  point to. It can be defined as

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

For example, again consider the article titled “Computer Science”. The out-links refer 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](https://en.wikipedia.org/wiki/Computer_Science)). The in-links and out-links have been diagrammatically demonstrated in Fig. 2.



Figure 2: In-links and 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, the 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 of top n terms as expansion terms.*
- *Re-weighting of expansion terms.*

### 3.2.1. Extraction of in-links

This step involves two sub-steps: (i) extraction of in-links and (ii) computation of term frequency ( $tf$ ) of the initial query terms. The in-links of an initial query term consist of the titles of all those Wikipedia articles that contain a hyperlink 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 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 term frequency of the word “Bird” and its synonyms obtained from WordNet in the article “Wings”.

### 3.2.2. Extraction of Out-links

Out-links of a query term are chosen 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 hyperlinks within the body of the article “Bird” are extracted as out-links.

### 3.2.3. 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. The semantic similarity has been calculated based on in-link scores. Let  $t$  be a query term and  $t_1$  be one of its candidate expansion terms. In reference

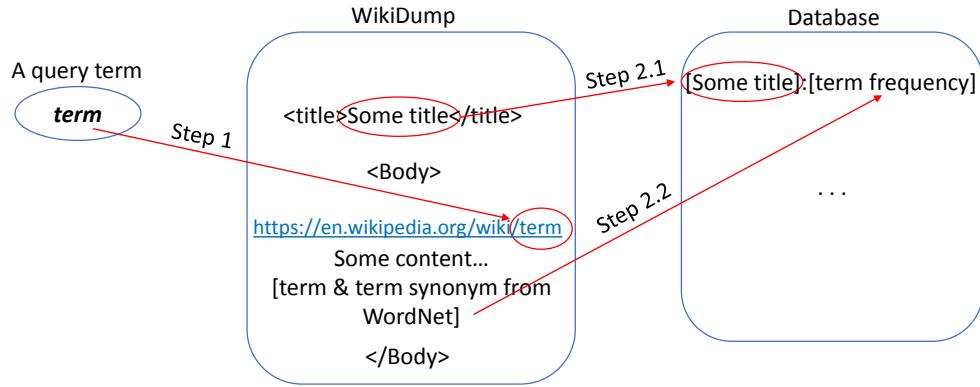


Figure 3: In-links Extraction

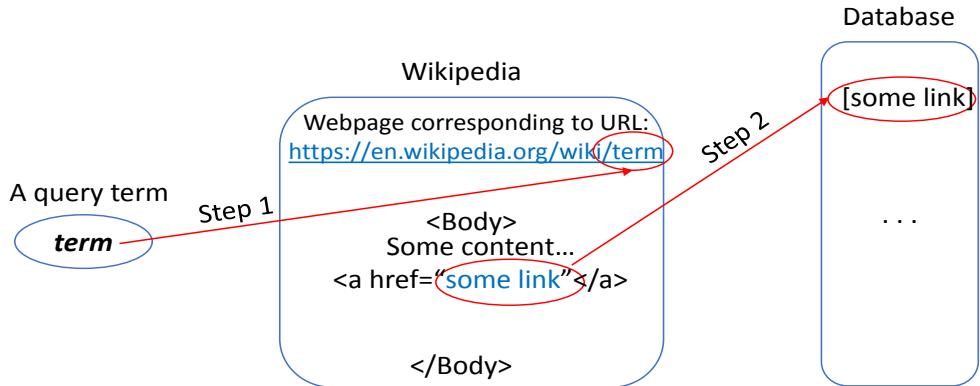


Figure 4: Out-links Extraction

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  $tf \cdot idf$  in IR and is calculated as follows:

$$Score(I(t_1)) = tf(t, t_1) \cdot idf(t_1, W_D) \quad (3)$$

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 follows:

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

where:

$N$  is the total number of articles in the Wikipedia dump, and

$|\{d \in W_D : t_1 \in d\}|$  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 the 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 a 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 hyperlinks of the query term article; the  $idf$  helps in removing these common hyperlinks that are 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 a correlation score (as described in Sec. 3.4). The top  $m$  terms chosen on the basis of the 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 are its individual terms 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 hyponym sets of the query term are considered as candidate expansion terms. Here synonyms and hyponyms are fetched at two levels, i.e., as shown in Fig. 5 for an initial query term  $Q_i$ , at level one its synonyms, denoted through  $x_i$ s, are considered, and, at level two, synonyms of  $x_i$ s, denoted through  $y_i$ s, are considered. The final synonym set used for QE is the union of level one and level two synonyms. Similarly, hyponyms are also fetched at two levels.

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

$$Score(t_1) = tf(t_1, t) \cdot idf(t_1, W_D) \quad (5)$$

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 the whole Wikipedia dump  $W_D$ .

$idf$  is calculated as given in Eq. 4.

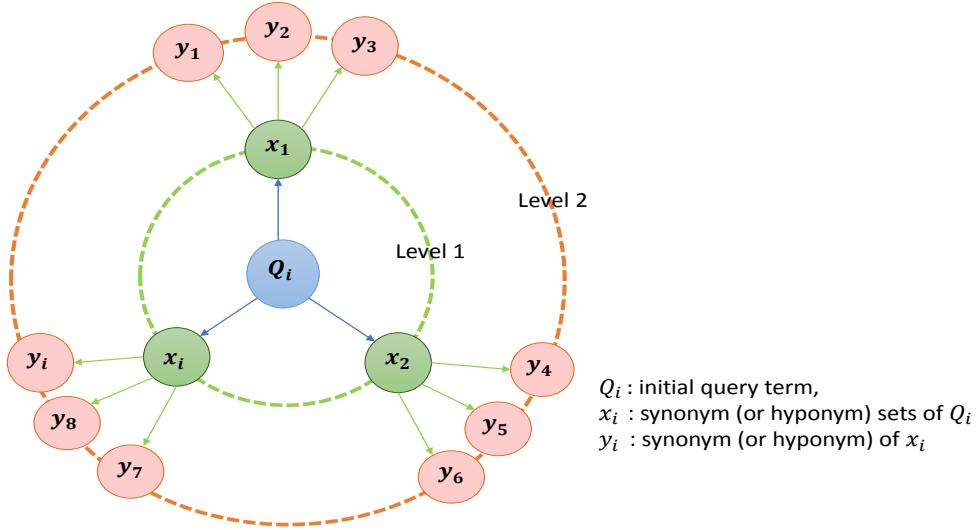


Figure 5: Initial query term and its two-level synonym (or hyponym) sets

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 the correlation score. Top  $m$  terms chosen on the basis of the 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  $tf \cdot idf$  score (for 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 [6]. To resolve this language ambiguity problem, we re-weight expanded terms using the correlation score [45]. 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:

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

where:

$c_{t,t_1}$  denotes the 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:

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

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. 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 phrases and individual words have been used for QE. We used the most recent Windows version of WordNet 2.1 to extract two levels of synsets terms and Wikipedia for in-links extraction for QE.

We use the Wikipedia 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 downloaded from <https://dumps.wikimedia.org/>. We downloaded the English Wikipedia dump titled "enwiki-20170101-pages-articles-multistream.xml" of January 2017.

We compared the performance of our query expansion technique with several existing weighting models and other related state-of-the-art techniques as described in Sec.4.2.

##### 4.1. Dataset

We use the well-known benchmark dataset Forum for Information Retrieval Evaluation (FIRE<sup>11</sup>) 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 the relevance of documents

---

<sup>11</sup><http://fire.irsi.res.in/fire/static/data>

to the corresponding topic(s). The FIRE dataset consists of a large collection of newswire articles from two sources, namely BDnews24<sup>12</sup> and The Telegraph<sup>13</sup>, 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<sup>14</sup> retrieval system for all our experimental evaluations. We used the title field of the topics in the FIRE dataset. For indexing the documents, first the stopwords were removed, and then Porter’s Stemmer was 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 [4], BM25 model of Okapi [38], Laplace’s law of succession I(n)L2 [20], Log-logistic DFR model LGD [12], Divergence from Independence model DPH [3], and Standard tf-idf model. The parameters for these models were set to the default values in TERRIER.

We evaluated the results on standard evaluation metrics: MAP (mean average precision), GM\_MAP (geometric mean average precision), F-Measure, 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 reported the percentage improvement in MAP over the baseline (non-expanded query) for each expansion method and other related methods. We also investigated the retrieval effectiveness of the proposed technique with a number of expansion terms.

## 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 three baseline approaches are: (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.

<sup>12</sup><https://bdnews24.com/>

<sup>13</sup><https://www.telegraphindia.com/>

<sup>14</sup><http://terrier.org/>

Table 2: Comparison of QE using Wikipedia alone on popular models with the 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						
Method	MAP	GM_MAP	P@10	P@20	P@30	#rel_ret
IFB2	0.3166 (↑14.5%)	0.2498 (↑30.99%)	0.4162 (↑13.72%)	0.3969 (↑11.49%)	0.3623 (↑5.94%)	2420 (↑3.86%)
I(n)L2	0.3317 (↑11.35%)	0.2628 (↑29.91%)	0.4425 (↑3.39%)	0.4012 (↑2.87%)	0.3892 (↑9.54%)	2432 (↑4.74%)
LGD	0.3248 (↑11.65%)	0.2535 (↑28.42%)	0.4432 (↑8.1%)	0.3901 (↑5.15%)	0.3639 (↑6.4%)	2428 (↑5.15%)
DPH	0.3291 (↑5.04%)	0.2598 (↑17.08%)	0.4667 (↑2.8%)	0.4127 (↑2.15%)	0.3783 (↑3.56%)	2423 (↑3.64%)
BM25	0.3304 (↑4.46%)	0.2501 (↑11.95%)	0.4723 (↑2.67%)	0.4044 (↑1.86%)	0.3717 (↑1.56%)	2421 (↑3.33%)
Tf-idf	0.3315 (↑4.15%)	0.2572 (↑13.75%)	0.4691 (↑2.87%)	0.4123 (↑2.82%)	0.3875 (↑4.53%)	2422 (↑3.5%)

Table 4 shows a 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.,  $\{0, 0.1, 0.2, 0.3, \dots, 1.0\}$ . These graphs are widely

Table 3: Comparison of QE using WordNet alone on popular models with the 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						
Method	MAP	GM_MAP	P@10	P@20	P@30	#rel_ret
IFB2	0.2901 (↑4.92%)	0.2113 (↑10.8%)	0.3817 (↑4.29%)	0.3693 (↑3.74%)	0.3521 (↑2.95%)	2361 (↑1.33%)
I(n)L2	0.3112 (↑4.46%)	0.2246 (↑11.02%)	0.4373 (↑2.17%)	0.3972 (↑1.85%)	0.3648 (↑2.67%)	2358 (↑1.55%)
LGD	0.3101 (↑6.6%)	0.2177 (↑10.28%)	0.4111 (↑0.27%)	0.3872 (↑4.37%)	0.3513 (↑2.72%)	2327 (↑0.78%)
DPH	0.3178 (↑1.43%)	0.2295 (↑3.42%)	0.4627 (↑1.92%)	0.4105 (↑1.61%)	0.3712 (↑1.62%)	2359 (↑0.89%)
BM25	0.3199 (↑1.14%)	0.2301 (↑3%)	0.4612 (↑0.26%)	0.3999 (↑0.73%)	0.3725 (↑1.78%)	2353 (↑0.43%)
Tf-idf	0.3203 (↑0.63%)	0.2312 (↑2.26%)	0.4597 (↑0.84%)	0.4098 (↑2.19%)	0.3827 (↑3.24%)	2345 (↑0.21%)

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 the P-R curve of the various weighting models have nearly the same retrieval result with or without QE respectively. Therefore we can say that for improving the information retrieval in QE, the 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 assigning a correlation score on the basis of the entire query.

Figure 7 compares the performance of the WWQE technique with P-R curve using popular

Table 4: Comparison of the proposed WWQE technique on popular models with the 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						
Method	MAP	GM_MAP	P@10	P@20	P@30	#rel_ret
IFB2	0.3439 (↑24.38%)	0.2835 (↑48.66%)	0.4660 (↑27.49%)	0.4400 (↑23.60%)	0.4040 (↑18.13%)	2554 (↑9.61%)
I(n)L2	0.3552 (↑19.23%)	0.2933 (↑44.98%)	0.4900 (↑14.48%)	0.4560 (↑16.92%)	0.4200 (↑18.21%)	2583 (↑11.24%)
LGD	0.3460 (↑18.94%)	0.2855 (↑44.63%)	0.4900 (↑19.51%)	0.4460 (↑20.21%)	0.4187 (↑22.43%)	2566 (↑11.13%)
DPH	0.3497 (↑11.62%)	0.2902 (↑30.78%)	0.4940 (↑8.81%)	0.4490 (↑11.14%)	0.4113 (↑12.59%)	2565 (↑9.71%)
BM25	0.3508 (↑10.91%)	0.2878 (↑28.83%)	0.5160 (↑12.17%)	0.4490 (↑13.10%)	0.4093 (↑11.83%)	2560 (↑9.26%)
Tf-idf	0.3521 (↑10.62%)	0.2896 (↑27.95%)	0.5100 (↑11.84%)	0.4520 (↑12.72%)	0.4120 (↑11.14%)	2561 (↑9.44%)

weighting models individually. Graphs in the figure show the improvement in retrieval results of the WWQE technique when compared with the initial unexpanded queries. As indicated in the graph legend, (Ex) denotes the performance with query expansion, while no parenthesis denotes unexpanded query. The P-R curves show the effectiveness of the proposed WWQE technique with all the popular weighting models. Of all the weighting models, the IFB2 weighting model provides the best retrieval performance with the proposed WWQE technique.

Figure 8 compares the WWQE technique in terms of precision, bpref and P@5 with various weighting models on the FIRE dataset in comparison to the unexpanded queries. Here, precision shows the ability of a system to present only relevant documents. P@5 measures the precision over the top 5 documents retrieved and bpref measures a preference relation about how many documents judged as relevant are ranked before the documents judged as irrelevant.

Figure 9 compares the WWQE technique in terms of MAP, bpref, F-Measure and P@5 with baseline (unexpanded), QE using WordNet alone, and QE using Wikipedia alone. IFB2 model

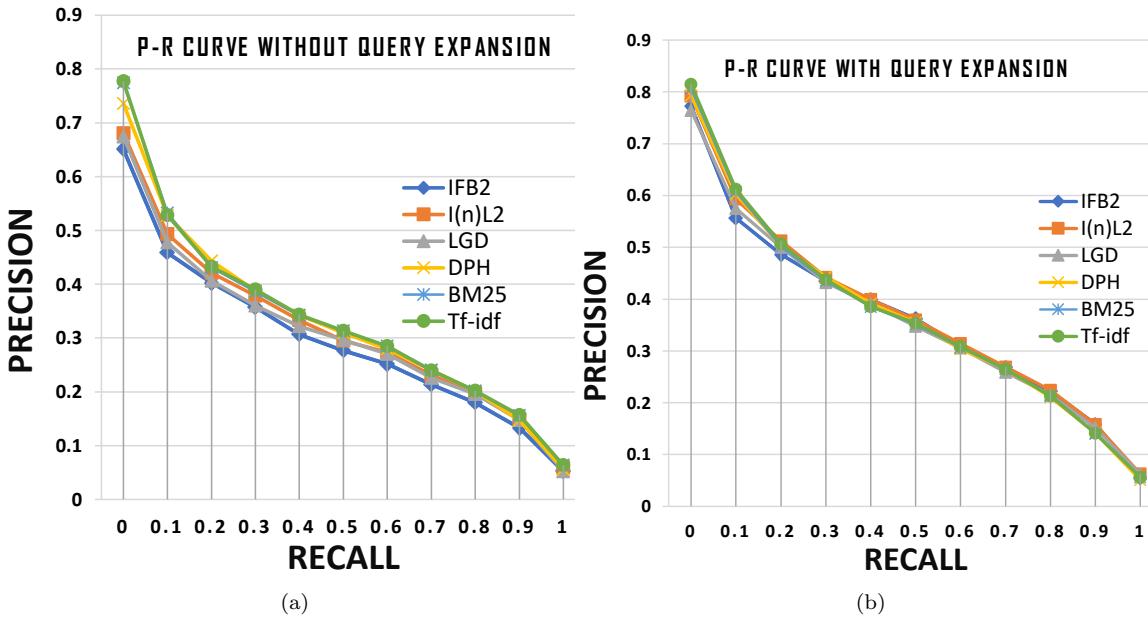


Figure 6: Comparative analysis of precision-recall curve of proposed QE technique with various weighting models on the FIRE dataset.

is used for term weighting for experimental evaluation. It can also be observed that the retrieval effectiveness of the proposed WWQE technique is superior to the QE using WordNet and Wikipedia alone.

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. Therefore, the proposed WWQE technique is effective in improving information retrieval results.

Table 5 presents a comparison of the proposed WWQE technique with Parapar et al. [36] and, Singh and Saran [41] models in terms of Mean Average Precision with the top 30 expansion terms on the FIRE dataset. We can observe that the retrieval effectiveness of the proposed WWQE technique is better than Parapar et al.'s and Singh and Saran's model. However, Parapar et al.'s and Singh and Saran's models perform well in comparison to the QE using WordNet and Wikipedia alone.

Parapar et al. [36] present an approach to minimize the number of non-relevant documents in the pseudo-relevant set to deal with the problem of unwanted documents in the pseudo-relevant collection. These unwanted documents adversely affect the selection of expansion terms. To automatically determine the number of documents to be selected for the pseudo-relevance set for each query, they studied the score distributions in the initial retrieval (i.e., documents retrieved in response to the initial query). The goal of their study was to come up with a threshold score to differentiate between relevant and non-relevant documents. Singh and Saran's [41] method uses a combination of pseudo-relevance feedback-based QE and multiple-term selection methods. They apply the Borda count rank-aggregating method to weigh expansion terms. Finally, the Word2vec

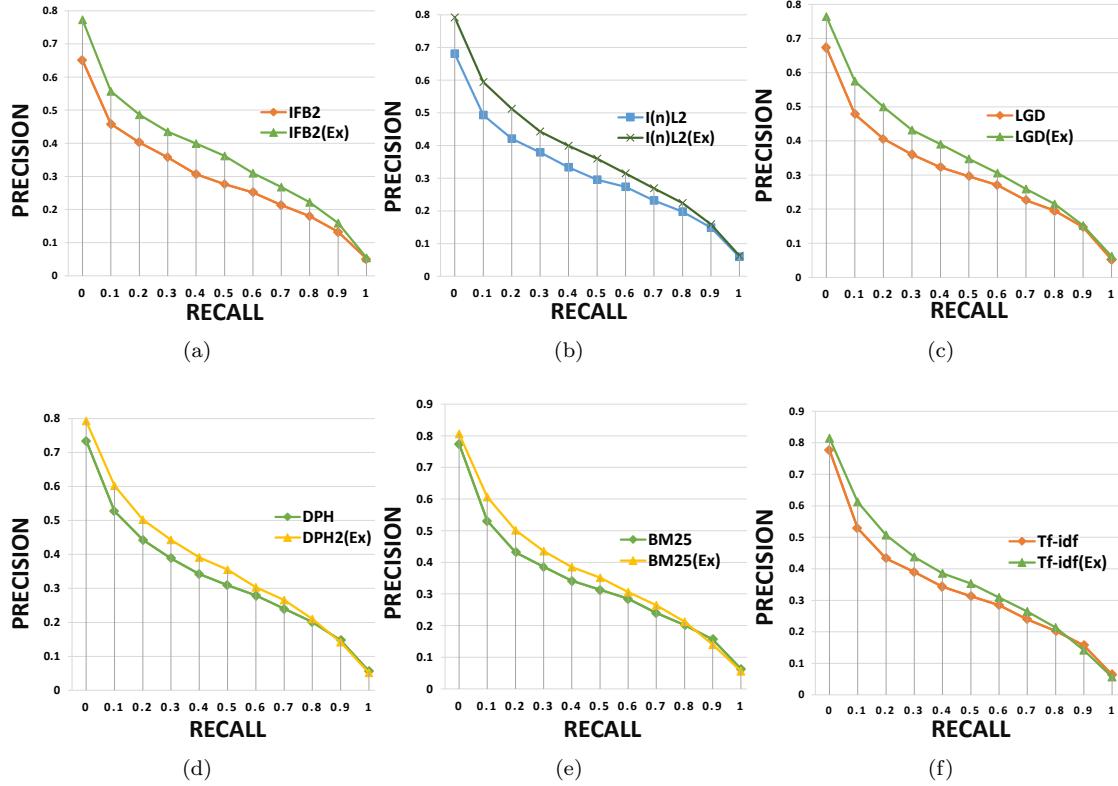


Figure 7: Comparative analysis of precision-recall curve of WWQE technique with popular weighting models individually on the FIRE dataset. In legend, (Ex) denotes the performance with query expansion, while no parenthesis denotes unexpanded query

approach is used to select semantically similar terms. The evaluation of their approach shows a significant improvement over individual term selection method.

Table 5: Comparative analysis of the proposed WWQE technique with Parapar et al.’s, and Singh & Saran’s models with respect to the MAP values

Data Set	Methods	MAP
FIRE	Baseline (IFB2)	0.2765
	WordNet model alone	0.2901 (4.92%)
	Wikipedia model alone	0.3166 (14.5%)
	Parapar et al. model [36]	0.3178 (14.74%)
	Singh and Saran model [41]	0.3217 (16.34%)
	WWQE (Proposed)	<b>0.3439 (24.38%)</b>

Figure 10 compares the WWQE technique in terms of the MAP, GM\_MAP, F-Measure and P@10 with baseline (IFB2), Parapar et al.’s, and Singh and Saran’s model. It can be clearly seen that the retrieval effectiveness of the proposed WWQE technique is better than the QE using Parapar et al.’s and Singh and Saran’s technique.

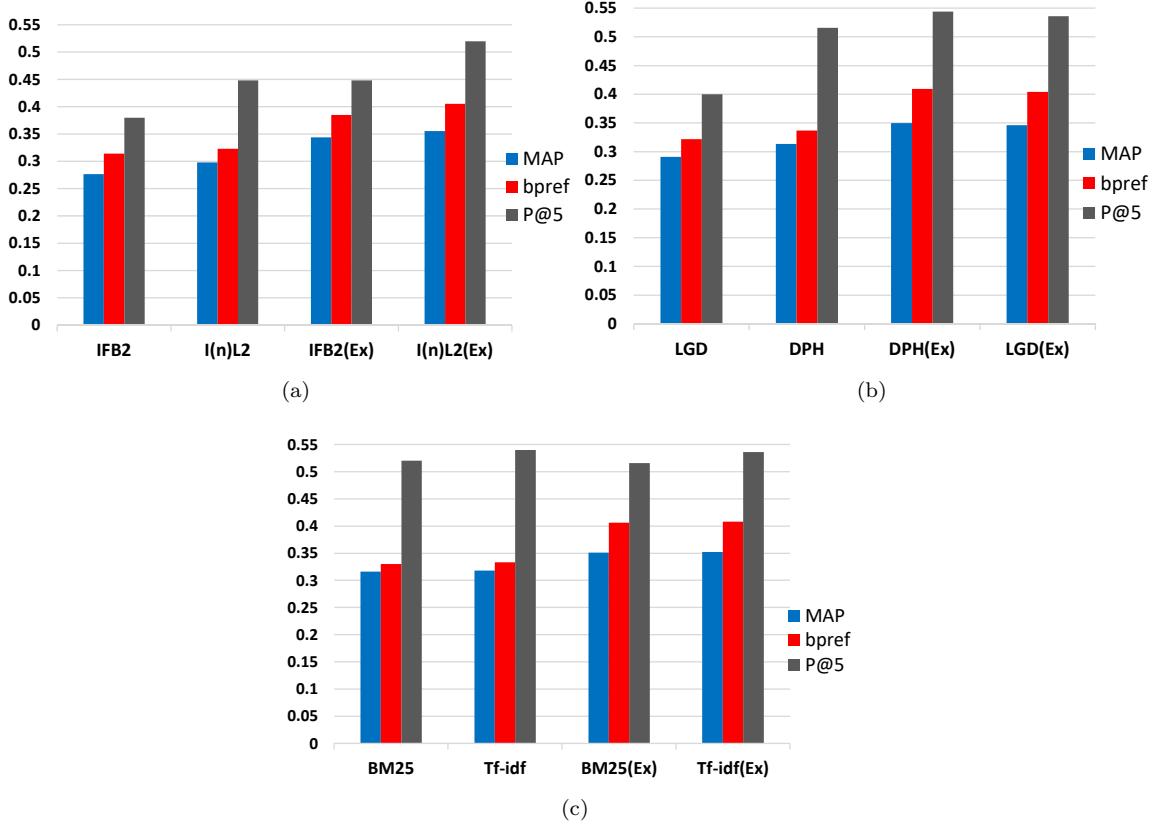


Figure 8: Comparative analysis of WWQE technique in terms of precision, bpref and P@5 with various weighting models on the FIRE dataset.

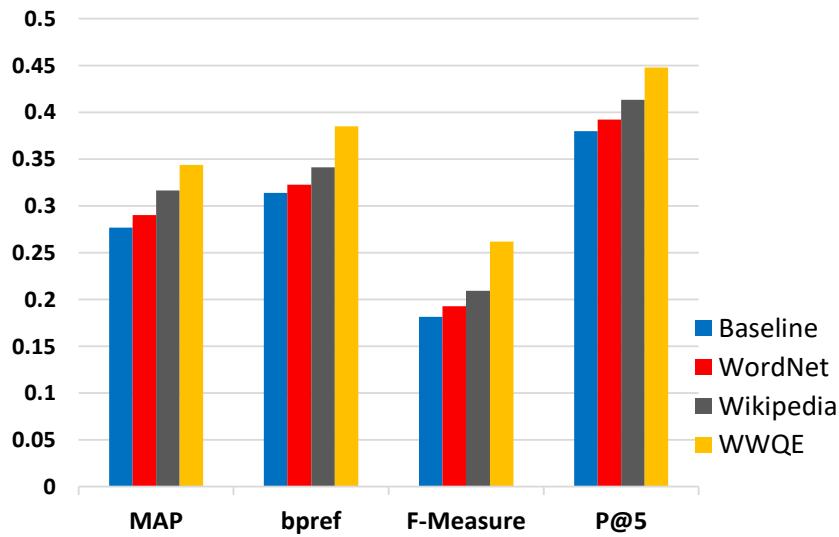


Figure 9: Comparative analysis of WWQE technique with baseline, WordNet and Wikipedia

### 5.1. Performance Variation with Number of Expansion Terms

There are different points of view on the number of expansion terms to be chosen; the number of expansion terms can vary from one-third of the expansion terms to all terms [5]. However, it might not be realistic to use all of the expansion terms, a small set of expansion terms is usually better

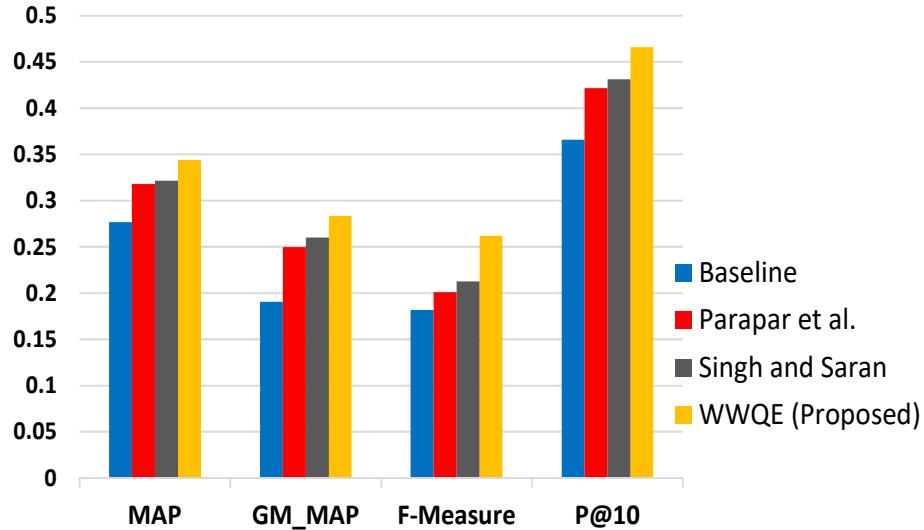


Figure 10: Comparative analysis of the WWQE technique with baseline, WordNet and Wikipedia

than a large set of expansion terms due to noise reduction [40]. A limited number of expansion terms may also be important to reduce the response time, especially for a large corpus. However, several studies observed that the number of expansion terms is of low relevance and it varies from query to query [9].

In our experimental results, we show the model performance with the top 30 expansion terms. The number 30 was chosen because it produced better retrieval performance compared to other integral values. We also did experiments by varying the expansion terms from 10 to 50 on our proposed model; Table 6 displays the corresponding results. The results show that the variation in performance is very limited with all weighting methods and for all selected sets of expansion terms. However, the retrieval performance of the system is adversely affected when we consider fewer than 20 expansion terms. As mentioned before, for our proposed model, 30 expansion terms seems to give the best results.

Table 6: Effect of the number of expansion terms on the performance (MAP) of the proposed WWQE technique on the FIRE dataset

Model Performance vs. Expansion terms						
Method	10	20	30	40	50	60
IFB2	0.2795	0.2907	<b>0.3439</b>	0.3401	0.3429	0.2781
I(n)L2	0.2997	0.2863	<b>0.3552</b>	0.3498	0.3426	0.3345
LGD	0.2981	0.3101	<b>0.3460</b>	0.3402	0.3421	0.3234
DPH	0.3301	<b>0.3499</b>	0.3497	0.3498	0.3365	0.3421
BM25	0.3251	0.3291	<b>0.3508</b>	0.3477	0.3379	0.3501
Tf-idf	0.3247	0.3365	0.3521	<b>0.3610</b>	0.3511	0.3367

Table 7 shows some examples of initial query and its expansion terms obtained with the pro-

posed approached.

Table 7: Expansion terms obtained with proposed approaches for selected queries on the FIRE dataset

Query ID	Original query	Expansion terms obtained with WordNet	Expansion terms obtained with Wikipedia	Expansion terms obtained with WWQE
126	Swine flu vaccine	Pig, influenza, swine influenza, immunogen, disease, antigen, illness, etc.	Human flu, Influenza, Influenza virus, H2N3, H1N2, Influenza pandemic, flu vaccine, etc.	Influenza viruses, pig, influenza, H1N1, H1N2, human flu, world health organisation, Infectious disease, etc.
128	Godhra train attack	Train, coach, transport, fire, firing, passenger train, human action, attack, etc.	Gujarat riots, Godhra train, Gujarat violence, Kar Sevak, Panchmahal district, etc.	Sabarmati Express, train, coach, Godhra, Gujarat, Hindu, Ayodhya, Babri Masjid, etc.
146	Ram Janmabhoomi verdict	Ram, verdict, act, judgement, group action, human activity, judicial decision, etc.	Ram Janmabhoomi, Ayodhya, Babri Masjid, Baqi Tashqandi, Barabanki district, Bharatiya Janata Party, etc.	Ayodhya dispute, Hindu, Rama, avatar, Hindutva, Ramayana, Babri Masjid, Ayodhya, Allahabad High Court, etc.
174	International economic slump	Global, economy, economic system, Market, drop-off, falloff, economic crisis, etc.	Economics, Demographic economics, Recession, Education economics, Markets, Public choice, Trade, etc.	Rescission, Macroeconomic, GDP, financial crisis, Stock market, Global recession, etc.

## 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. The proposed method employs a two-level strategy to select terms from WordNet. First, it fetches synsets of the initial query terms. Then, it extracts synsets of these synsets. In order to score the expansion terms on Wikipedia, we proposed a new weighting score named as an in-link score. We employed a tf-idf-based scoring system to assign a score to expansion terms extracted from WordNet. After

assigning a score to individual query terms, we further re-weighted the selected expansion terms using the 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 two methods individually and with the other related state-of-the-art methods. This article also investigates the retrieval performance of the proposed technique with varying numbers of expansion terms. The results on the basis of several evaluation metrics and popular weighting models on the FIRE dataset demonstrate the effectiveness of our proposed QE technique in the field of information retrieval.

### Acknowledgments

Akshay Deepak has been awarded Young Faculty Research Fellowship (YFRF) of Visvesvaraya PhD Programme of Ministry of Electronics & Information Technology, MeitY, Government of India. In this regard, he would like to acknowledge that this publication is an outcome of the R&D work undertaken in the project under the Visvesvaraya PhD Scheme of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

### References

#### References

- [1] Al-Shboul, B. and Myaeng, S.-H. (2014). Wikipedia-based query phrase expansion in patent class search. *Information retrieval*, 17(5-6):430–451.
- [2] ALMasri, M., Berrut, C., and Chevallet, J.-P. (2013). Wikipedia-based semantic query enrichment. In *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*, pages 5–8. ACM.
- [3] Amati, G., Amodeo, G., Bianchi, M., Gaibisso, C., and Gambosi, G. (2008). Fub, iasi-cnr and university of tor vergata at trec 2008 blog track. Technical report, FONDAZIONE UGO BORDONI ROME (ITALY).
- [4] Amati, G. and Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389.
- [5] Azad, H. K. and Deepak, A. (2017). Query expansion techniques for information retrieval: a survey. *arXiv preprint arXiv:1708.00247*.

- [6] Bai, J., Nie, J.-Y., Cao, G., and Bouchard, H. (2007). Using query contexts in information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 15–22. ACM.
- [7] Bai, J., Song, D., Bruza, P., Nie, J.-Y., and Cao, G. (2005). Query expansion using term relationships in language models for information retrieval. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 688–695. ACM.
- [8] Bhogal, J., Macfarlane, A., and Smith, P. (2007). A review of ontology based query expansion. *Information processing & management*, 43(4):866–886.
- [9] Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., and Li, H. (2008). 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*, pages 875–883. ACM.
- [10] Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1.
- [11] Chakraborty, G. and Pagolu, M. K. (2014). Analysis of unstructured data: Applications of text analytics and sentiment mining. In *SAS global forum*, pages 1288–2014.
- [12] Clinchant, S. and Gaussier, E. (2010). Information-based models for ad hoc ir. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 234–241. ACM.
- [13] Collins-Thompson, K. and Callan, J. (2005). Query expansion using random walk models. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 704–711. ACM.
- [14] Dalton, J., Dietz, L., and Allan, J. (2014). Entity query feature expansion using knowledge base links. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 365–374. ACM.
- [15] Dang, V. and Croft, B. W. (2010). Query reformulation using anchor text. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 41–50. ACM.
- [16] Dong, S., Liu, J., Liu, Y., Zeng, L., Xu, C., and Zhou, T. (2018). Clustering based on grid and local density with priority-based expansion for multi-density data. *Information Sciences*, 468:103–116.
- [17] Elsas, J. L., Arguello, J., Callan, J., and Carbonell, J. G. (2008). 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*, pages 347–354. ACM.

- [18] Fan, H., Ma, Z., Wang, D., and Liu, J. (2018). Handling distributed xml queries over large xml data based on mapreduce framework. *Information Sciences*, 453:1–20.
- [19] Fang, H. (2008). A re-examination of query expansion using lexical resources. *proceedings of ACL-08: HLT*, pages 139–147.
- [20] Good, I. (1965). The estimation of probabilities: An essay on modern bayesian methods, pp. xi-xii.
- [21] Ji, Z., Ma, Y., Pang, Y., and Li, X. (2019). Query-aware sparse coding for web multi-video summarization. *Information Sciences*, 478:152–166.
- [22] Lee, J., Yu, I., Park, J., and Kim, D.-W. (2019). Memetic feature selection for multilabel text categorization using label frequency difference. *Information Sciences*, 485:263–280.
- [23] Lemos, O. A., de Paula, A. C., Zanichelli, F. C., and Lopes, C. V. (2014). Thesaurus-based automatic query expansion for interface-driven code search. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, pages 212–221. ACM.
- [24] Li, Y., Li, G., Li, J., and Yao, K. (2018). Skqai: A novel air index for spatial keyword query processing in road networks. *Information Sciences*, 430:17–38.
- [25] Li, Y., Luk, W. P. R., Ho, K. S. E., and Chung, F. L. K. (2007). 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*, pages 797–798. ACM.
- [26] Liu, J., Liu, C., and Huang, Y. (2017). Multi-granularity sequence labeling model for acronym expansion identification. *Information Sciences*, 378:462–474.
- [27] Liu, S., Liu, F., Yu, C., and Meng, W. (2004). 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*, pages 266–272. ACM.
- [28] Liu, S., Wu, J., Feng, L., Qiao, H., Liu, Y., Luo, W., and Wang, W. (2018). Perceptual uniform descriptor and ranking on manifold for image retrieval. *Information Sciences*, 424:235–249.
- [29] Lu, M., Sun, X., Wang, S., Lo, D., and Duan, Y. (2015). Query expansion via wordnet for effective code search. In *Software Analysis, Evolution and Reengineering (SANER), 2015 IEEE 22nd International Conference on*, pages 545–549. IEEE.
- [30] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.
- [31] Maron, M. E. and Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, 7(3):216–244.

- [32] Merigó, J. M., Pedrycz, W., Weber, R., and de la Sotta, C. (2018). Fifty years of information sciences: A bibliometric overview. *Information Sciences*, 432:245–268.
- [33] Nakamura, T. A., Calais, P. H., de Castro Reis, D., and Lemos, A. P. (2019). An anatomy for neural search engines. *Information Sciences*, 480:339–353.
- [34] Pal, D., Mitra, M., and Bhattacharya, S. (2015). Exploring query categorisation for query expansion: A study. *arXiv preprint arXiv:1509.05567*.
- [35] Pal, D., Mitra, M., and Datta, K. (2014). Improving query expansion using wordnet. *Journal of the Association for Information Science and Technology*, 65(12):2469–2478.
- [36] Parapar, J., Presedo-Quindimil, M. A., and Barreiro, A. (2014). Score distributions for pseudo relevance feedback. *Information Sciences*, 273:171–181.
- [37] Pedrycz, W., Krawczak, M., and Zadrożny, S. (2018). Computational intelligence techniques for decision support, data mining and information searching. *Information Sciences*, 460-461:374–376.
- [38] Robertson, S. E., Walker, S., Beaulieu, M., Gatford, M., and Payne, A. (1996). Okapi at trec-4. *Nist Special Publication Sp*, pages 73–96.
- [39] Rocchio, J. J. (1971). Relevance feedback in information retrieval.
- [40] Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American society for information science*, 41(4):288–297.
- [41] Singh, J. and Sharan, A. (2016). Relevance feedback-based query expansion model using ranks combining and word2vec approach. *IETE Journal of Research*, 62(5):591–604.
- [42] Smeaton, A. F., Kelledy, F., and O'Donnell, R. (1995). Trec-4 experiments at dublin city university: Thresholding posting lists, query expansion with wordnet and pos tagging of spanish. *Harman [6]*, pages 373–389.
- [43] Stein, R. A., Jaques, P. A., and Valiati, J. F. (2019). An analysis of hierarchical text classification using word embeddings. *Information Sciences*, 471:216–232.
- [44] Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In *SIGIR94*, pages 61–69. Springer.
- [45] Xu, J. and Croft, W. B. (1996). 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*, pages 4–11. ACM.
- [46] Xu, Y., Jones, G. J., and Wang, B. (2009). Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 59–66. ACM.

- [47] Zhang, J., Deng, B., and Li, X. (2009). Concept based query expansion using wordnet. In *Proceedings of the 2009 international e-conference on advanced science and technology*, pages 52–55. IEEE Computer Society.
- [48] Zhou, D., Zhao, W., Wu, X., Lawless, S., and Liu, J. (2018). An iterative method for personalized results adaptation in cross-language search. *Information Sciences*, 430:200–215.