Query Expansion Techniques for Information Retrieval: a Survey

Article · August 2017

CITATIONS
10

READS
1,823

2 authors:

Hiteshwar Kumar Azad
National Institute of Technology Patna
9 PUBLICATIONS **52** CITATIONS
SEE PROFILE

Akshay Deepak
National Institute of Technology Patna
33 PUBLICATIONS **59** CITATIONS
SEE PROFILE

Noname manuscript No.
(will be inserted by the editor)

Query Expansion Techniques for Information Retrieval: a Survey

Hiteshwar Kumar Azad · Akshay Deepak

Received: date / Accepted: date

Abstract With the ever increasing size of web, relevant information extraction on the Internet with a query formed by a few keywords has become a big challenge. To overcome this, query expansion (QE) plays a crucial role in improving the Internet searches, where the user's initial query is reformulated to a new query by adding new meaningful terms with similar significance. QE – as part of information retrieval (IR) – has long attracted researchers' attention. It has also become very influential in the field of personalized social document, Question Answering over Linked Data (QALD), and, Text Retrieval Conference (TREC) and REAL sets. This paper surveys QE techniques in IR from 1960 to 2017 with respect to core techniques, data sources used, weighting and ranking methodologies, user participation and applications (of QE techniques) – bringing out similarities and differences.

Keywords Query expansion · Query reformulation · Information retrieval · Internet search

1 Introduction

There is huge amount of data available on Internet and it is growing exponentially. This unconstrained information-growth has not been accompanied by an analogous expansion of approaches for extracting relevant information [185]. Often, a web-search does not yield relevant results. There are multiple reasons for this. First, the keywords submitted by the user can be related to multiple topics; the search results are not focused on the topic of interest. Second, the query can be too short to express appropriately what the user is looking for. This can happen simply as a matter of habit (average size of a web search is 2.4 words[1] [249]). Third, the user is often not sure about what he is looking for until he sees the results. Fourth, even if the user knows what he is searching for, he does not know how to formulate the appropriate query (navigational queries are exception to this [49]). QE plays an important part in fetching relevant results in the above cases.

Most web queries fall under the three fundamental categories [49,137] :

- *Informational Queries.* Queries that cover a broad topic (e.g., *India* or *journals*) for which there may be thousands of relevant results.
- *Navigational Queries.* Queries that looking for specific website or URL (e.g., *ISRO*).
- *Transactional Queries.* Queries that demonstrate the user intent to execute a specific activity (e.g., downloading papers or buying books).

Hiteshwar Kumar Azad
National Institute of Technology Patna
Tel.: +91-9852118731
E-mail: hiteshwar.cse15@nitp.ac.in

Akshay Deepak
National Institute of Technology Patna
E-mail: akshayd@nitp.ac.in

[1] https://www.statista.com/statistics/269740/number-of-search-terms-in-internet-research-in-the-us/

Currently, user queries are mostly processed using indexes and ontologies, which work on exact matches and are hidden from the users. This leads to the problem of term mismatch: user and search index don't use the same terms. Also known as the vocabulary problem [97]; it results from the combination of synonymy and polysemy. Synonymy refers to multiple words with common meaning, e.g., such as "buy" and "purchase". Polysemy refers to words with multiple meanings, e.g., "mouse" (a computer device or an animal). Synonymous and polysemous words are hindrance in retrieving relevant information; they reduce recall and precision rates.

Various techniques have been proposed to address vocabulary problem, including relevance feedback, interactive query filtration, corpus depended knowledge models, corpus independent knowledge models, search result clustering, and word sense disambiguation. Almost all popular techniques expand the initial query by adding new related terms. This can also involve selective retention of terms from the original query. The expanded/reformulated query is then used to retrieve more relevant results. The whole process is called Query expansion (QE).

Query expansion has a long history in literature. It was first implied in 1960 by Moron et al. [179] as a technique for literature indexing and searching in a mechanized library system. Rocchio [229] brought QE to spotlight; the author used "relevance feedback" and characterized it in vector space model. The idea behind relevance feedback is to incorporate user's feedback in the retrieval process so as to improve the final result. In particular, the user gives feedback on the retrieved documents in response to the initial query by indicating relevance of the results. Rocchio's work was further extended and applied in techniques such as collection-based term co-occurrence [131, 219], cluster-based information retrieval [127,188], comparative analysis of term distribution [212, 283,259] and automatic text processing [234,232,233].

This above was before the search engine era, when search-retrieval was done on a small amount of data with short queries and satisfactory result were obtained. Search engines were introduced in 1990s and a vast amount of data was suddenly published on the web, which has continued to grow at an exponential rate since then. Users continued to fire short queries for web searches. While the recall rate suddenly increased, there was a loss in precision [235,109]. This called for modernization of QE techniques to deal with Internet-data.

As per a recent report [2] [3], the most frequent queries consist of one, two or three words only (see Fig. 1) – the same as seventeen year ago as reported in reference [154]. While the number of query terms have remained few, the number of web pages have increased exponentially. This has increased the ambiguity – caused due to multiple meanings/senses of query term(s) or vocabulary mismatch problem – in finding relevant pages. Hence, the importance of query expansion (QE) techniques has also increased in resolving the vocabulary mismatch problem.

Recently, QE has come to spotlight because a lot of researchers are using the QE technique for working on personalized social bookmarking services [102,36,47], Question Answering over Linked Data (QALD)[4] [257] and in Text Retrieval Conference (TREC)[5]. They are also used heavily in web, desktop and email searches [204]. 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.

However, there are also drawbacks of QE techniques, e.g., there is a computational cost associated with the application of QE techniques. In case of Internet searches, where quick response time is a must, the computational cost associated with application of QE techniques prohibits their use [123] in part of entirety. Another drawback is that sometimes it can fail to establish relationship between a word in the corpus with those which are used in different communities, e.g., "senior citizen" and "elderly" [101]. Another issue is that query expansion may hurt the retrieval effectiveness for some queries [66,174].

Few surveys have been done in past on QE techniques. In 2007, Bhogal et al. [34] reviewed the query expansion using an ontology, which are domain specific. Such techniques have also been described in book [177]. Carpineto et al. [58] (published in year 2012) reviewed the major QE

[2] https://www.statista.com/statistics/269740/number-of-search-terms-in-internet-research-in-the-us/
[3] http://www.keyworddiscovery.com/keyword-stats.html
[4] http://qald.sebastianwalter.org/
[5] http://trec.nist.gov/
[6] https://wordnet.princeton.edu/
[7] http://conceptnet5.media.mit.edu/
[8] http://lucene.apache.org/
[9] https://enterprise.google.com/search/

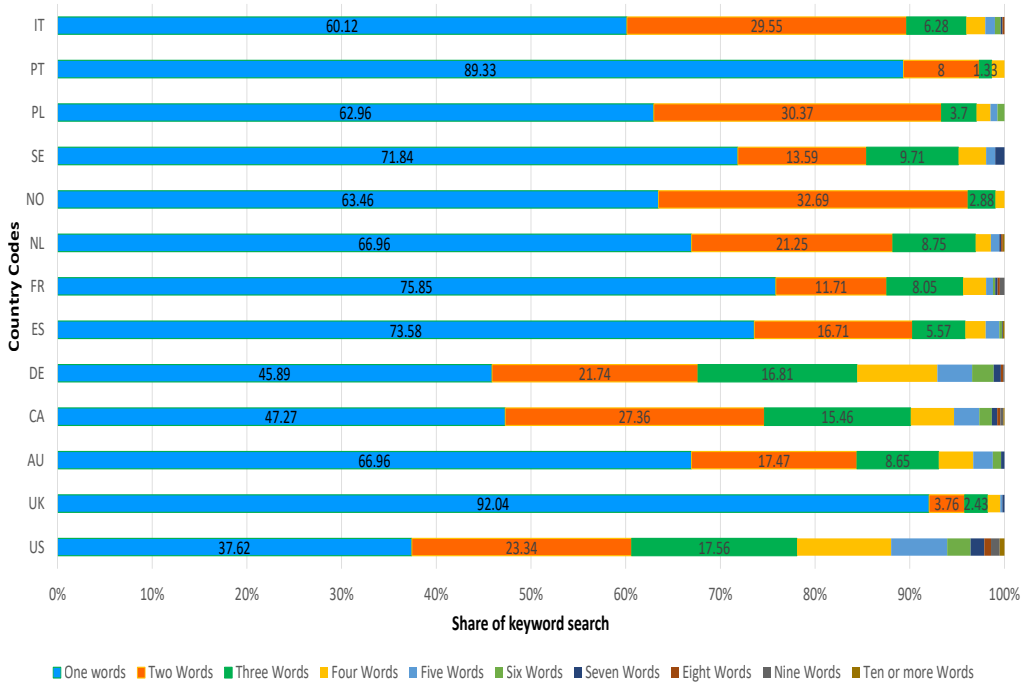

Fig. 1: Query size searched by different countries

techniques, Data sources and features in an information retrieval system. After this, we could not find any major review covering recent progress in QE techniques. However, reference [58] covers only automatic query expansion techniques and does not include recent research on personalized social documents, term weighting and ranking methods and categorization of several data sources. In contrast, this survey – in addition to covering recent research in QE techniques – also covers research on automatic, manual and interactive QE techniques. This paper discusses QE techniques from four key aspects: (i) data sources, (ii) applications, (iii) working methodology and (iv) core approaches (see Fig. 2).

The rest of the article is organized as follows. Section 2 defines QE and discuss the importance and application of QE. It also briefly discusses several applications of QE including recent literatures. Section 3 describes the working methodology of query expansion and outlines the main steps. Section 4 classifies the existing approaches on the basis of data sources. Finally, Section 5 discuss recent trend in literature and concludes our work.

2 Query Expansion

Query expansion reformulates user's original query to enhance the information retrieval effectiveness. Let a user query consist of n terms, Q = $\{t_1, t_2, ..., t_i, t_{i+1}, ..., t_n\}$. The reformulated query can have two components: addition of new terms T'=$\{t'_1, t'_2, ..., t'_m\}$ from the data source(s) D and removal of stop words T''= $\{t_{i+1}, t_{i+2}, ..., t_n\}$. The reformulated query can be represented as:

$$\begin{aligned} Q_{exp} &= (Q - T'') \cup T' \\ &= \{t_1, t_2, ..., t_i, t'_1, t'_2, ..., t'_m\} \end{aligned} \tag{1}$$

In the above definition, the key aspect of QE is set T': set of new meaningful terms added to user's original query to retrieve more relevant documents and reduce ambiguity. Karovetz et al. [146] reported that this set T' computed on basis of term similarity, and without changing the concept, increases recall rate in query results. Hence, computation of set T' and choice of data sources D is the key aspect of research.

In regard to automation and end user involvement [88], QE techniques can be classified as follows:

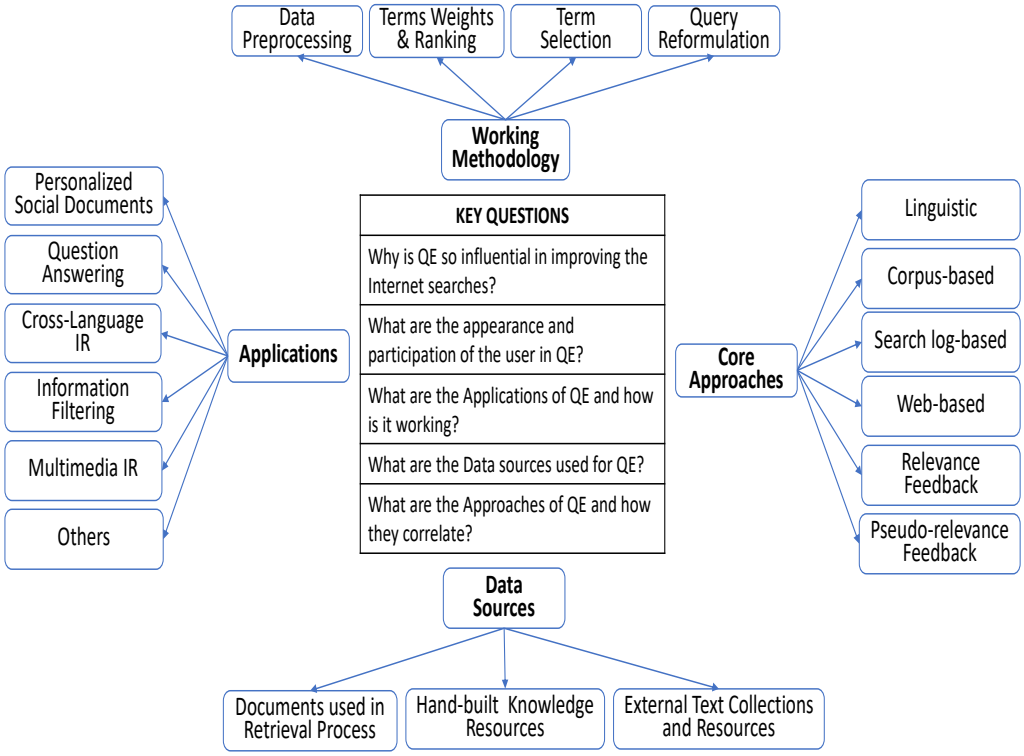

Fig. 2: Survey overview

– *Manual Query Expansion.* Here, user manually reformulates query.
– *Automatic Query Expansion.* Here, system automatically reformulates query – without any user intervention. Both the technique to compute set T' and choice of data sources D is incorporated into system's intelligence.
– *Interactive Query Expansion.* Here, query reformulation happens as a result of joint cooperation between the system and user. It is a human-in-the-loop approach, where the system returns search results on an automatically reformulated query and the users choses meaningful results among them. Based on user's preference, the system further reformulates query and retrieves results. The process continues till the user is satisfied with the search results.

2.1 Importance of Query Expansion

One of the major importance of the QE is that it enhances the chance to retrieve the relevant information on the Internet, which is not retrieved otherwise using the original query. While this improves the recall ratio, attempt to retrieve a lot of relevant documents adversely affects precision. Many times user's original query is not sufficient to retrieve the information user intends or is looking for. In this situation, QE plays a crucial role in improving the Internet searches. For example, if the user's original query is *Novel*, it is not understandable what the user wants: the user may be searching for fictitious narrative book or the user may be interested in something new or unusual. So QE expands the original query *"Novel"* to *"Novel book" "Book" "New" "Novel approach"*. This new query retrieves the documents which have both types of meaning. This technique has been used hugely for search operations in various commercial domains (such as education, hospitality in medical science, economics, experimental research [58]) where the primary intension is to retrieve entire relevant documents related to a particular concern. It has been experimentally observed that during query expansion, attempt to increase recall rate adversely affects precision ratio and vice versa [110,112,202]. The main reason behind the loss in precision is that the relevant documents retrieved in response to the user's initial query may rank lower in ranking after query expansion. For improvement of retrieval precision, expanded query can also use Boolean operators (AND,OR) or PARAGRAPH operator [189] to transform expanded query to Boolean query [207,141], which

is eventually submitted for retrieval. For example, let the expanded query (from Eq 1) be T_{exp}= {t_1,t_2, . . .,t_i, t'_1, t'_2,...,t'_m}. The expanded Boolean query can be B_{query}= {t_1 AND t_2 AND...AND t_i AND t'_1 AND t'_2 AND...AND t'_m}, or, B_{query}= {t_1 OR t_2 OR...OR t_i OR t'_1 OR t'_2 OR...OR t'_m}, or, a combination of OR and AND operators. A common issue with AND operator is that it improves precision but reduces recall rate, whereas, OR operator reduce precision but improve recall rate (e.g., [256]). Reference [141] proposes a novel Boolean query suggestion technique, where Boolean queries are produced by exploiting decision trees learned from pseudo-labeled documents and ranks produced queries using query quality predictors. The authors compared this technique to the recent query expansion techniques and experimentally demonstrated its superiority. XML queries can also used for improving the precision in IR system for enhancing the Internet searches (e.g., [136,63,135]). Improving precision in IR system through query expansion using web pages has been proposed in references [74,75,293]. Here, query expansion happens based on collection of important words in related web pages. Another set techniques for the same task expanded queries using query concept [214,93,115,77]. Here, expansion happens based on similar meaning of query terms.

However, for considering the joint evaluation of precision and recall rate in query expansion, several experimental study agree that retrieval quality of query results is enhanced by ten percentage or more with expansion of the user query (e.g., [236,278,59,157,89,221,69,53]). Such reports support the effectiveness of query expansion techniques in information retrieval system. Some recent studies have shown that query expansion can also improve the precision by disambiguating the user query (e.g., [251,22,291,281]). Table 1 show the efforts of several researcher for improving the precision & recall rate.

Table 1: Summary of Techniques used for Improving Precision & Recall rate

Expansion Techniques	**Publications**
Boolean query	Pane and Myers 2000 [207], Moldovan and Mihalcea 2000 [189], Kim et al. 2011 [141]
XML queries	Kamps et al. 2006 [136], Chu-Carroll et al. 2006 [63], Junedi et al. 2012 [135]
Collection of the top terms within the web pages	Cui et al. 2002 [74], Cui et al. 2003 [75], Zhou et al. 2012 [293]
Query concepts	Qiu and Frei 1993 [214], Fonseca et al. 2005 [93], Hsu et al. 2008 [115], Bouchoucha et al. 2013 [48]
Query Disambiguation	Stokoe et al. 2003 [251], Bai et al. 2005 [22], Zhong and Ng 2012 [291], Yao et al. 2015 [281]

2.2 Application of Query Expansion

Beyond the key area of information retrieval, there are other recent applications where QE technique has proved beneficial. We discuss some of such applications next.

2.2.1 Personalized Social Documents

In recent years social tagging systems have achieved popularity by being used in sharing, tagging, commenting, rating, etc., of multi-media contents. Every user wants to find relevant information according to their interests and commitments. This has generated need of a query expansion framework that is based on social bookmarking and tagging systems, which enhance document representation.

Reference [29] discusses exploiting the social relations for query expansion using users, tags and documents. Reference [37] uses the social tagging and bookmarking in the query expansion for personalized web searches. The experimental results show effective matching of user's interests and search results. Reference [46] uses the combination of social proximity and semantic similarity for personalized social query expansion, which is based on similar terms that are mostly used by a given user and their social relatives. Reference [293] proposed a query expansion technique that is based on distinctive user profiles in which terms are extracted from both the annotations and

resources the user has created and opted (also used in [44]). Many other works (e.g., [45,107, 193]) discuss the query expansion and social personalized ranking in the context of personalized social documents. Recently an article [47] proposed a technique PerSaDoR (Personalized social document representation), where for indexing and modeling user's activities in a social tagging system is used, and, for query expansion social annotations are used. A more recent work in personalized IR [12] uses word embedding for query expansion, where experimental evaluation was done on the collection of CLEF Social Book Search 2016[10].The main motive of this paper is to address the following questions: (1) "How to use the word embedding for QE in the context of social collection?", and, (2) "How to use the word embedding to personalize query expansion?" Reference [295] personalizes query expansion using enriched user profiles on the web; the user profiles have were created using external corpus (folksonomy data). They also proposed a model to enhance the user profile. This model integrates the word embeddings with topic models in two groups of pseudo-relevant documents such as user annotations and documents from the external corpus.

2.2.2 Question Answering

Question Answering (QA) has become a very influential research area in the field of information retrieval system. The main objective of QA is to grant a quick answer in response to a user's query. Here the focus is to keep the answer concise rather than retrieving all relevant documents. The system accepts as input natural language questions, such as "Who is the first nation in the world to enter Mars orbit in first attempt?", instead of a set of terms. Recently search engines have also started using the QA system to provide answer to such types of questions. However, for ranking answers of such questions, the main challenges in QE is mismatch problem, which arises due to mismatch between the expression in question and answer texts [164].

To overcome the mismatch problem and improving the document retrieval in QA system, many articles have been published. In 2004, reference [4] presented a common approach for query expansion using FAQ data; the same approach was also followed in reference [248]. Reference [218] presents a technique to expand the user's original query in QA system using Statistical Machine Translation (SMT), which linked the verbal gap between user's questions and system's answers. Reference [35] ranked the retrieved documents in QA using social media search and web search. Other works [208,167,60,190] use social network for improving the retrieval performance in QA. Reference [271] expands short queries by mining the user intent from three different sources including community question answering (CQA) archive, query logs and web search results. Currently query expansion on question answering over linked open data (QALD) has gain much attention in the area of natural language processing. An article [239] has proposed an approach for expansion of the original query on linked data using linguistic and semantic features, where linguistic features are extracted from WordNet and semantic features are extracted from Linked Open Data (LOD)[11] cloud. The evaluation was carried out on a training dataset extracted from the QALD question answering benchmark. /the experimental results show a considerable improvement in precision and recall values over baseline approaches.

2.2.3 Cross-Language Information Retrieval (CLIR)

It is the part of IR which retrieves the information present in languages different from the user's query language. For example a user's query is in Hindi but his retrieved relevant information can be in English. Over the past few years CLIR is achieving more attention due to the Popularity of CLEF[12] and TREC which are held annually for promoting the research in the area of IR.

Traditionally there are three main approaches to CLIR : query translation with machine translation techniques [215], parallel or comparable corpora-based techniques [240] and machine readable bilingual dictionaries [23]. The main issue with the traditional CLIR is untranslatable query terms, phrase translation, inflected term, and uncertainty in language translation between source and target languages [210]. To overcome this translation error, a popular approach is to use query

[10] http://social-book-search.humanities.uva.nl
[11] http://lod-cloud.net/
[12] http://www.clef-initiative.eu/

expansion [24,198]. It gives better output even in case of no translation error due to the use of statistical semantic similarity among the terms [2,143]. To counter the errors in automated machine translation in case of cross-language queries, reference [98] uses the linguistic resources for query expansion. Query expansion can be applied at various points in the translation process: before or after translation or both. It has been shown that application at prior translation gives better result in comparison to application at post translation, however, the application at both steps has given best results [25,26,181,160]. For improving the query expansion in CLIR reference [54] combines the dictionary translation and the co-occurrence term relations into Markov Chain (MC) models – defined as a directed graph where query translation is formulated as a random walk in MC models. Recent work [292,294] used query expansion techniques for personalize CLIR based on user's historical usage information. Experimental results show that personalized approaches work better than non-personalized approaches in CLIR. Reference [33] presents a technique to translate the query from Hindi to English CLIR using word embedding.

2.2.4 Information Filtering

Information filtering (IF) is a method to eliminate inessential information from the entire dataset and deliver the relevant results to the end user. Information filtering is widely used in various domains such as searching Internet, e-mail, e-commerce, multimedia distributed system, blogs (for survey see [108]). There are two basic approaches for IF: content-based filtering and collaborative filtering. Reference [28] discusses the relationship between IR and IF and establishes that IF is a special kind of IR, and has the same research challenges/outcomes. They have a common goal to provide the relevant information to the user but from different aspects. Reference [108] gives a brief overview of IF and discuss the difference between IR and IF with respect to research issues. For improving the relevancy of results obtained after IF, several QE approaches have been published. Relevance feedback techniques expanded query can reflect the user's interest and need well [8]. Reference [91] combines user's query with system's master query [91] for improving results. Other techniques include using user's profile [284], using geographical query footprint [96], using correlated keywords [296], using the links and anchor text in Wikipedia [14], using text classification in twitter messages [250] and using the patterns of online users behavior [287] [100]. A more recent work [273] reformulated the query using user-item co-clustering method for improving the Collaborative Filtering. Another reference [285] reorganizes query using DBpedia[13] corpus and ClueWeb09[14] corpus for efficient Boolean IF.

2.2.5 Multimedia Information Retrieval

Multimedia information retrieval (MIR) deals with searching and extracting the semantic information from multimedia documents (such as audio, video and image) (see [161] for review). Most of the MIR systems typically rely on text-based search on multimedia documents such as title, captions, anchor text, annotations and surrounding html or xml depiction. This approach can fail when metadata is absent or when the metadata cannot exactly describe the actual multimedia content. Hence, query expansion plays a crucial role in extracting the most relevant multimedia data.

Audio retrieval deals with searching audio files in large collections of audio files. The retrieved files should be similar to the audio query, which is natural language. The search analyzes the actual contents of the audio rather than the metadata such as keywords, tags, and/or descriptions associated with the audio. For searching spoken audio, a common approach is to do text search on transcription of the audio file. However, transcription is obtained automatically by speech translation software and, hence, contain errors. In such a case, expanding the transcription by adding related words greatly improves retrieval effectiveness [244]. However, for text document retrieval, benefits of such document expansion are limited [265]. A study [134] shows that query expansion can improve the average precision by 17 percent in audio retrieval. Reference [27] follows QE technique based on semantic similarity in audio IR. Reference [253] compares the language dependent and language independent query through examples and concludes that language dependent setup

[13] http://wiki.dbpedia.org/
[14] http://lemurproject.org/clueweb09/

provides better results in spoken term detection. Recently, reference [140] presented a query expansion method for social audio contents, where query expansion approach uses three speech segments: semantic, window and discourse-based segments.

In video retrieval, queries and documents have both visual as well as textual aspect. The expanded text queries are matched with manually established text descriptions of the visual concepts. Reference [194] expands text query using lexical, statistical and content-based approaches for visual query expansion. Reference [254] expanded the query using the corpus of natural language description based on exact evaluation of system performance. A more recent work [255] uses the meta synopsis for video indexing; the meta synopsis contains vital information for retrieving relevant videos.

In image retrieval, a common approach to retrieve relevant images is querying using textures, shapes, color and visual aspect that match with image descriptions in the database (e.g., for review on image retrieval [162] [81]). Reference [148] presents two query expansion approaches: intra expansion (expanded query is obtained from existing query features) and inter expansion (expanded query is obtained from search results). Reference [117] uses the query logs data for generic web image searches. Reference [43] uses multitag for image retrieval, whereas [166] retrieves images using query adaptive hashing method. Reference [275] presents a contextual query expansion technique to overcome (a) the semantic gap of visual vocabulary quantization, and, (b) performance and storage loss due to query expansion in image retrieval .

2.2.6 Other Applications

Some other recent applications of query expansion are plagiarism detection [197], event search [87, 15, 41], text classification [262], patent retrieval [175, 176, 261], in dynamic process in IoT [120, 121], classification of e-commerce [126], biomedical IR [1], enterprise search [170], code search [199] and twitter search [147, 297].

Table 2 summarizes some of the prominent and recent applications of QE in literature based on above discussion.

3 Query Expansion Working Methodology

The process of generating query expansion consists of mainly four steps: preprocessing of data sources, term weights and ranking, term selection and query reformulation (see Fig. 3). Each step has been discussed in the following sections.

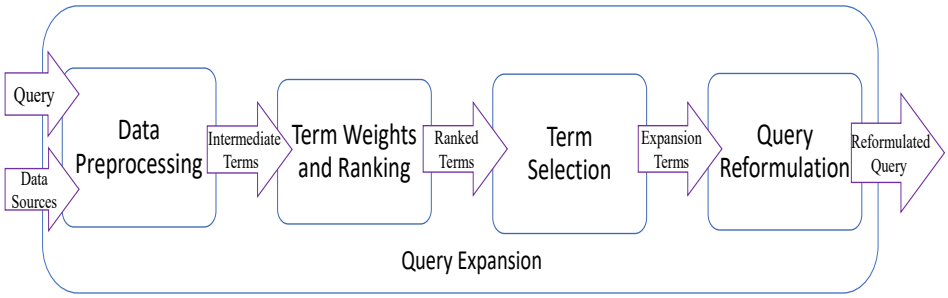

Fig. 3: Query expansion working model

3.1 Preprocessing of Data Sources

Preprocessing of a data source is depends upon the data sources and approaches used for query expansion, instead of user's query. The primary goal of this step (preprocessing of data source) is to extract a set of terms form data sources that meaningfully augment user's original query. It consists of the following four sub-steps:

Table 2: Summary of Research in Applications of Query Expansion

Research Area	Data Sources	Applications	Publications
Personalized social document	Social annotations, user logs, social tag and bookmarking, social proximity and semantic similarity, word embedding, social context	Enhance document's representation and grant a personalized representation of documents to the user	Zhou et al. 2017 [295], Bouadjenek et al. 2016 [47], Amer et al. 2016 [12], Mulhem et al. 2016 [193], Hahm et al. 2014 [107], Bouadjenek et al. 2013 [44], Zhou et al. 2012 [293], Bouadjenek et al. 2011 [46], Biancalana and Micarelli 2009 [37]
Question Answering	FAQs, QA pairs, Social network, WordNet, LOD cloud, community question answering (CQA) archive, query logs and web search	Responsind to user's query with quick concise answers rather than retunring all relevant documents	Molino et al. 2016 [190], Cavalin et al. 2016 [60], Liu et al. 2015 [167], Shekarpour et al. 2013 [239], Panovich et al. 2012 [208],Bian et al. 2008 [35], Riezler et al. 2007 [218]
Cross-Language Information Retrieval	User logs, word embeddings, dictionary translations and co-occurrence terms, linguistic resources	Retrieving information written in a language different from user's query language	Zhou et al. 2016 [294], Bhattacharya et al. 2016 [33], Zhou et al. 2015 [292], Gaillard et al. 2010 [98],Cao et al. 2007 [54], Kraaij et al. 2003 [143]
Information Filtering	User profile, user log, anchor text, Wikipedia, DBpedia corpus, twitter messages	Searching results on Internet, e-mail, e-commerce and multimedia distributed system	Zervakis et al. 2016 [285], Wu et al. 2016 [273], Gao et al. 2015 [100], Zhang and Zeng 2012 [287], Arguello et al. 2008 [14], Fu et al. 2005 [96], Yu et al. 2004 [284]
Multimedia Information Retrieval	Title, captions, anchor text, annotations, meta synopsis, query logs, multitag, corpus of natural language and surrounding html or xml depiction	Searching and extracting semantic information from multimedia documents (audio, video and image) such as audio retrieval, video retrieval and image retrieval	Khwileh and Jones 2016 [140], Thomas et al. 2016 [255], Li et al. 2016 [162], Xie et al. 2014 [275], Liu et al. 2013 [166], Tejedor et al. 2012 [253], Tellex et al. 2010 [254], Kuo et al. 2009 [148], Wei and Croft 2007 [265]
Others	Word embeddings, CLEF- IP patent data, Top documents, Wikipedia, DBpedia, TREC collection, Genomic data sets, top tweets, etc.	Text classification, patent retrieval, plagiarism detection, dynamic process in IoT, twitter search, biomedical IR, code search, event search, enterprise search	Wang et al. 2016 [262], Wang and Lin 2016 [261], Nawab et al. 2016 [197], Huber et al. 2016 [120], Zingla et al. 2016 [297], Abdulla et al. 2016 [1], Nie et al. 2016 [199], Atefeh et al. 2015 [15], Liu et al. 2014 [170]

1. Text extraction from data sources (extraction of the whole texts from the specific data source used for query expansion)
2. Tokenization (process of splitting the stream of texts into words)
3. Stop word removal (removal of frequently used words e.g., articles, adjective, prepositions, etc.)
4. Word stemming (process for reduction of derived or inflected words to their base word)

A lot of data sources have been used for QE in literature. All such sources can be classified into three classes: (i) documents used in retrieval process, (ii) hand-built knowledge resources, and (iii) external text collections and resources.

3.1.1 Documents Used in Retrieval Process

In the beginning of seventies, addition of clustered term into the initial query started playing a crucial role in query expansion (e.g.,[127,188,268]). Researchers assume that the set of similar words that appear frequently in documents belongs to the same subject and similar documents

form a cluster [209]. Two types of clustering have been discussed in documents retrieval system: clustering of terms and clustering of documents [268]. Reference [214] is a well known corpus-based expansion technique [214] that uses similarity thesaurus for expanding the original query. A similarity thesaurus is a collection of documents based on specific domain knowledge, where each term is expressed as a weighted document vector. Some other works use collection-based data sources for query expansion (e.g., [131,16,73,277,101,57,22]). Recently, article [290] used four corpora as data sources (including one industry and three academic corpus) and presented a Two-stage Feature Selection framework (TFS) for query expansion known as Supervised Query Expansion (SQE). The first stage is Adaptive Expansion Decision (AED), which predicts whether a query is suitable for SQE or not. For unsuitable queries, SQE is skipped with no term features being extracted at all, so that computation time is reduced. For suitable queries, the second stage conducts Cost Constrained Feature Selection (CCFS), which chooses a subset of effective yet inexpensive features for supervised learning. A drawback of corpus specific query expansion is that they fail to establish relationship between a word in the corpus and those which are used in different communities, e.g., "senior citizen" and "elderly" [101].

While researchers agree that the addition of selective terms improve the retrieval effectiveness, there are differing point of views on the number of selective terms to be added; ranging form one third of the terms [222], five to ten terms [11,61], 20-30 terms [109,290], 30-40 terms [201], few hundreds terms [31,269] to 350-530 terms for each query [52]. The source of these terms came can be the top retrieved documents or well known relevant documents. Further, it improves retrieval effectiveness by 7% to 25% [52].

3.1.2 Hand-built Knowledge Resources

The main goal of the hand-built knowledge resources is to extract knowledge from textual hand-built data sources such as dictionaries, thesaurus, ontologies and LOD cloud. Thesaurus-based query expansion can be either automatic or hand-built. One of the famous hand-built thesaurus is WordNet [186]. Reference [260] utilized the WordNet to expand the original query with semantically similar terms called synsets. It was observed that retrieval effectiveness was improved significantly for unstructured query, while only marginal improvement was observed for structured queries. Some other articles also use the WordNet to expanded the original query (e.g., [245] uses synonyms of the initial query and assigns half weight,[169] uses word sense, [103] uses semantic similarity, [288] uses concepts and [206] uses semantic relations form WordNet). Reference [206] proposes a new and effective way of using WordNet for query expansion, where candidate expansion terms (CET) are selected from a set of pseudo-relevant documents and the usefulness of these terms determined by considering multiple sources of information. The semantic relation between expanded terms and the query terms is determined using WordNet. Reference [159] presents an automatic query expansion (AQE) approach that uses word relations to increase the chances of finding relevant code. For data source, it uses a thesaurus containing only software-related word relations and WordNet for expanding the user's query. Similarly, reference [114] uses ConceptNet [168] (have higher concept diversity) and WordNet (have higher discrimination ability) as the data sources for expanding user query. ConceptNet is a relational semantic network that helps to understand the common sense knowledge of text written by users. Recently a number of researchers used ConceptNet as the data source for query expansion (e.g., [115,142,48,13]). Reference [48] uses ConceptNet for query expansion and proposes a QE technique known as Maximal Marginal Relevance-based Expansion (MMRE). This technique selects expansion terms that are closely related to the initial query, but different from the previously selected expansion terms. Then, the top N expansion terms having the highest MMRE scores are selected. Reference [17] uses LOD cloud for keyword mapping and exploits the graph structure within Linked Data to determine relations between resources that are useful to discover, or directly express semantic similarity. Utilization of data sources as knowledge bases in information retrieval is still an open problem because most of the prior research focuses on the construction of knowledge bases rather than their utilization techniques. Presently, knowledge bases (describes entities, their attributes, and their relationships to other entities) is an influential data sources for query expansion. Recent reference [276] uses the knowledge bases such as freebase (large public knowledge base that contains semi-structured information about real world entities and their facts) for improving the query expansion. For selection of the expansion term, reference [276] developed two methods: (1) utilization of tf.idf based Pseudo Relevance Feedback (PRF) on

linked objects' descriptions, and, (2) utilization of Freebase's entity categories which grants an ontology tree that illustrate entities at several levels of abstraction.

However, reference [113] used thesaurus relationship for query expansion in the UMLS Metathesaurus and reported that nearly all types of query expansion reduce the recall and precision based on retrieval effectiveness. Not surprisingly, in their result, only 38.6% of the queries with synonym expansion and up to 29.7% of the queries with hierarchical expansion showed significant improvement in retrieval performance. Primarily, there are three limitations in hand-built knowledge resources: they are commonly domain specific, usually do not contain proper noun and they have to be kept up to date. Experiments with query expansion using hand-built knowledge resources did not show consistent improvements in retrieval effectiveness. It does not improve well formulated user queries, but significantly improve the retrieval effectiveness of badly constructed queries.

3.1.3 External Text Collections and Resources

Text collection used in retrieval process (such as the WWW, Wikipedia) is the most common and effective data sources for query expansion. In such cases, query expansion approaches shows overall better performance in comparison to all three data sources. Some data sources under this category need preprocessing procedures for text collection. For example, reference [144,79] uses the anchor texts as the data sources; it parses hyperlinks to extract data from anchor tags. Further, additional steps need to be carried out such as stop word removal and word stemming. Their experimental results also suggests that anchor texts can also be used to improve the traditional QE based on query logs. Query logs is another data source for query expansion, where user's queries are expanded using correlation between query terms and document terms determined using user logs (e.g., [266, 75]). Some researchers refer to query logs as user logs since it is derived from historical records of user queries registered in the query logs of search engine (e.g., [74,38,18,282]). Reference [282] expresses the search engine query log as a bipartite graph, where query nodes are connected to the URL nodes by click edges; it reported improvement of retrieval effectiveness by more than 10%. Reference [264] uses web corpus and training data as data sources, and then, extracts query using search logs. Most of the search engines and surveyed papers using QE are based on the query logs. However, customized search systems in the Internet search, enterprise search, personalized search (such as desktop or email search) or for infrequent queries, query logs are either not available or the past user's queries are not sufficient to describe the information needed. To overcome this limitation reference [32] proposed a probabilistic model that provides the expansion terms from the corpus without using query logs.

A good number of published works have used hybrid data sources (combining two or more data sources) for query expansion. For example, reference [67] uses WordNet as data source, an external corpus, Krovetz stemmer and top retrieve documents. Recently, reference [205] used data source based on term distributions (based on Kullback-Leibler Divergence (KLD) and Bose-Einstein statistics (Bo1)) and term association (Local Context Analysis (LCA) and Relevance-based Language Model (RM3)) methods for query expansion. The experimental result demonstrated that the combined method gives better result in comparison to the each individual method. Other research works based on hybrid resources are [111,157,271,77]. Recently, Wikipedia and DBpedia are using widely as data sources for query expansion (e.g., [163,14,279,3,9,13,106]). Reference [163] performed an investigation using Wikipedia and retrieved all articles corresponding to the original query as a source of the expansion terms for pseudo relevance feedback (PRF). 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 [279] 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 and cover a narrow topic) and, (3) broader queries (queries having neither ambiguous nor specific entity). They consolidate the expansion terms into the original query and evaluate these techniques using language modeling information retrieval. Reference [9] uses Wikipedia for semantic enrichment of short queries based on in-link and out-link articles. Reference [77] propose 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), including structured attributes and text. The main motive for linking

entities to knowledge bases is to improve the understanding and representation of text documents and queries. In reference [13], document collection and external resources (encyclopedias such as DBpedia and knowledge bases such as ConceptNet) are the data sources for query expansion. For selecting the expansion terms, term graphs have been constructed using information theoretic measures based on co-occurrence between each pair of terms in the vocabulary of the document collection. Today, word embedding techniques are widely used for query expansion. Recently, reference [230] proposed word embedding framework (based on distributed neural language model word2vec). Based on the framework it extracted similar terms to a query using K-nearest neighbor approach. The experimental study was done on standard TREC ad-hoc data; it showed considerable improvement over the classic term overlapping-based retrieval approach. It should also be noticed that word2vec based query expansion methods perform more or less the same with and without any feedback information. Some other articles using word embedding techniques are [84, 149]. Reference [84] presented a QE technique based on locally-trained word embedding (such as word2vec and GloVe) for ad hoc information retrieval. It also used local embeddings that capture the nuances of topic-specific language and are better than global embeddings. It also suggested that embeddings be learned on topically-constrained corpora, instead of large topically-unconstrained corpora. In a query-specific manner, their experimental results suggested towards adopting the local embeddings instead of global embedding for formers potentially superior representation. Similarly reference [149] proposed a QE technique based on word embeddings that uses Word2Vecs Continuous Bag-of-Words (CBOW) approach [184]; CBOW represents terms in a vector space based on their co-occurrence in text windows. It also presents a technique for integrating the terms selected using word embeddings with an effective pseudo relevance feedback method.

Recently, fuzzy logic based expansion techniques have also become popular. References [243, 242] used a fuzzy logic-based query expansion technique, and, top-retrieved documents (using pseudo-relevance feedback) as data sources. Here, each expansion term (obtained from top retrieved documents) is given a relevance score using fuzzy rules. The relevance scores of the expanded terms are summed up to infer the high fuzzy weights for selecting expansion terms.

Table 3 summarizes the classification of Data Sources used in QE in literature based on above discussion.

3.2 Weighting and Ranking of Query Expansion Terms

In this step of QE, weights and ranks have been assigned to query expansion terms [see Fig. 3]. The input of this step is the user's query as well as the texts extracted from the data sources in the first step. Assigned weights denote relevancy of terms in the expanded query and are further used in ranking retrieved documents based on relevancy. There are many techniques for weighting and ranking of query expansion terms. Reference [58] classifies the techniques into four categories on the basis of relationship between the query terms and the expansion features:

- *One-to-One Association.* Such as WordNet to find synonyms and similar terms for the query terms.
- *One-to-Many Association.* Correlates one query term to many expanded query terms.
- *Feature Distribution of Top Ranked Documents.* Deals with the top retrieved documents from the initial query and considers the top weighted terms from these documents.
- *Query Language Modeling.* Constructs a statistical model for the query and choses the expansion terms having highest probability.

The first two approaches can also be considered as local techniques. These are based on association hypothesis projected by reference [220]: "If an index term is good at discriminating relevant from non-relevant documents, then any closely associated index term is also likely to be good at this." This hypothesis was mainly motivated by reference [178]. Reference [220] outlines this concept as "to enlarge the initial request by using additional index terms which have a similar or related meaning to those of the given request." The above approaches have been discussed next.

Table 3: Summary of Research in Classification of Data Sources used in Query Expansion

Type of Data Sources	Data Sources	Term Extraction Methodology	Publications
Documents Used in Retrieval Process	Clustered terms	Clustering of terms and documents from sets of similar objects	Jardine and Rijsbergen 1971 [127], Minker et al. 1972 [188], Willett 1988 [268]
	Corpus or Collection based data sources	Terms collection from specific domain knowledge	Jones 1971 [131], Attar and Fraenkel 1977 [16], Peat and Willett 1991 [209], Crouch and Yang 1992 [73], Qiu and Frei 1993 [214], Xu and Croft 1996[277], Gauch et al. 1999 [101], Carpineto et al. 2001 [57], Bai et al. 2005 [22]
Hand Built Knowledge Resources	WordNet & Thesaurus	Word sense and synset	Miller et al. 1990 [186], Voorhees 1994 [260], Smeaton et al. 1995 [245], Liu et al. 2004 [169], Gong et al. 2006 [103], Zhang et al. 2009 [288], Pal et al. 2014 [206]
	ConceptNet & Knowledge bases	Common sense knowledge and Freebase	Liu and Singh 2004 [168], Hsu et al. 2006 [114], Hsu et al. 2008 [115], Kotov and Zhai 2012 [142], Bouadjenek et al. 2013 [45], Anand and Kotov 2015 [13]
External Text Collections and Resources	Anchor texts	Adjacent terms in anchor text or text extraction from anchor tags	Kraft and Zien 2004 [144], Dang and Croft 2010 [79]
	Query logs or User logs	Historical records of user queries registered in the query logs of search engine	Wen et al. 2002 [266], Cui et al. 2003 [75], Billerbeck et al. 2003 [38], Baeza-Yates et al. 2004 [18], Yin et al. 2009 [282], Hua et al. 2013 [117]
	Wikipedia or DBpedia	articles, titles & hyper links	Li et al. 2007 [163], Arguello et al. 2008 [14], Xu et al. 2009 [279], Aggarwal and Buitelaar 2012 [3], ALMasri et al. 2013 [9], Al-Shboul and Myaeng 2014 [7], Anand and Kotov 2015 [13], Guisado-Gamez et al. 2016 [106]
	External corpus	Nearby terms in word embedding framework	Roy et al. 2016 [230], Diaz et al. 2016 [84], Kuzi et al. 2016 [149]
	Top-ranked documents & Hybrid sources	All terms in top retrieved documents	Collins-Thompson and Callan 2005 [67], He and Ounis 2007 [111], Lee et al. 2008 [157], Pal et al. 2013[205], Wu et al. 2014 [271], Dalton et al. 2014 [77], Singh and Sharan 2016 [243]

3.2.1 One-to-One Association

This is the basic approach for weighting and ranking the expansion terms based on one-to-one association between the query terms and expansion terms. Here, each expansion term is connected to a individual query term and weights are assign for each query terms using several techniques.

One of the most influential approaches to establish one-to-one association is to use linguistic associations namely stemming algorithm to minimize the inflected word (plural forms, tenses of verbs or derived forms) from its word stem. For example, based on Porter's stemming algorithm [211], the words "stems", "stemmed", "stemming" and "stemmer" would be reduced to the root word "stem." Another typical linguistic approach is the use of thesaurus. One of the most famous thesaurus is WordNet [260]; each query term is mapped to its synonyms and similar set of words – obtained from WordNet – in the expanded query. For example, if we consider word "java" as noun in WordNet, there are three synsets each having a specific sense: for location (as an island), food (as coffee), and computer science (as a programming language). The same approach has been done using ConceptNet [114] to find related concepts of user's queries for query expansion. For example, word "file" having related concept as "folder of document", "record", "computer", "drawer", "smooth rough edge", "hand tool", etc. Then, each expanded term is assigned a similarity

score based on their similarity with the query term. Only terms with high scores are retained in expanded query. The natural concept regarding term similarity is that two terms are semantically similar, if both terms are in the same document. Similarly two documents are similar if both are having the same terms. There are several approaches to determine term similarity.

Path length-based measure determines the term similarity between the synsets(sense) – obtained from WordNet – based on path length of linked synsets. Generally path length-based measures included two similarity measurement techniques: Shortest path similarity (derived from [216]) and Wu & Palmer [274]. Let the given terms be s_1 and s_2, and let len_s denote the length of the shortest path between s_1 and s_2 in WordNet. The Shortest path similarity score is define as:

$$Sim_{Path}(s_1, s_2) = \frac{1}{len_s} \tag{2}$$

Path length between members of the same synset is considered to be 1, hence, the maximum similarity score will be 1.

The Wu & Palmer (WP) similarity score is defined as

$$Sim_{WP}(s_1, s_2) = \frac{2 \,.\, d(LCS)}{d(s_1) + d(s_2)} \tag{3}$$

where, $d(LCS)$ is the depth of Least common sub-sumer(LCS)(the closest common ancestor node of two synsets) and $d(s_1)$ & $d(s_2)$ are the depth of senses s_1 and s_2 from the root node R in WordNet (see Figure 4). Similarity score of WP varies from 0 to 1 ($0 < Sim_{WP} \leq 1$).

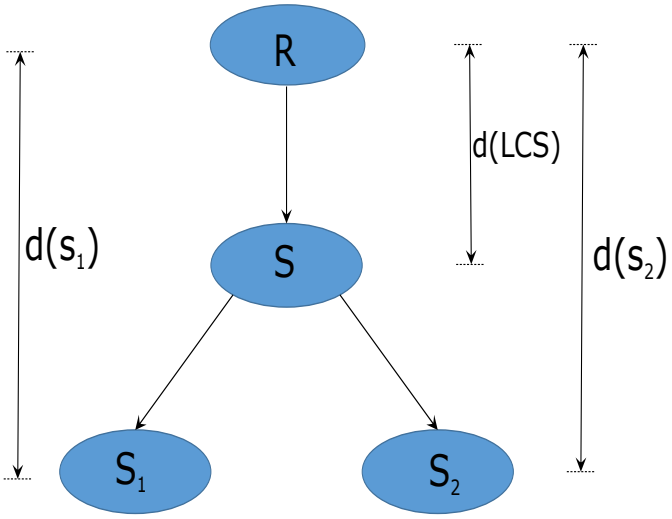

Fig. 4: Example of taxonomy hierarchy in WordNet

Other approaches like Jaccard coefficient and Dice coefficient are also used widely for similarity measurement. The Jaccard coefficient [124] is described as:

$$Sim_{Jaccard}(s_1, s_2) = \frac{df_{s_1 \wedge s_2}}{df_{s_1 \vee s_2}} \tag{4}$$

where, $df_{s_1 \wedge s_2}$ denotes the frequency of documents containing both s_1 and s_2, and $df_{s_1 \vee s_2}$ denotes the frequency of documents containing at least s_1 or s_2.

The Dice coefficient [85] is described as

$$Sim_{Dice}(s_1, s_2) = \frac{2 \,.\, df_{s_1 \wedge s_2}}{df_{s_1} + df s_2} \tag{5}$$

where df_{s_1} and $df s_2$ denote the frequency of documents containing s_1 and s_2 respectively.

A more generic approach term-document matrix M is a two dimensional matrix, whose rows represent the terms and columns represent the documents. Cell $M_{t,d}$ contains value $w_{t,d}$, where $w_{t,d}$ denotes weight of term t in document d. Correlation matrix $C = MM^T$, where each cell c_{s_1,s_2} denotes correlation (similarity) score between terms s_1 and s_2 is described as

$$c_{s_1,s_2} = \sum_{d_j} w_{s_1,j} \,.\, w_{s_2,j} \tag{6}$$

The cosine similarity measure, Sim_{cosine} is defined as normalization of the above correlation factors:

$$Sim_{cosine} = \frac{c_{s_1,s_2}}{\sqrt{\sum_{d_j} w_{s_1,s_1}^2 \cdot \sum_{d_j} w_{s_2,s_2}^2}} \tag{7}$$

Normalization is done to account for relative frequency of terms.

As we see, using equation (6), we can create a set of conceptually different term-to-term correlation method by varying how to select the set of documents and the weighting function. Although, calculating co-occurrence of all terms present in the document is easy, it does not consider relative position of terms in a document. For example, two terms that co-occur in the same sentence are more correlated than when they occur in distant parts of the document.

A more exhaustive measurement technique for term co-occurrence that includes term dependency is mutual information [64]:

$$I_{s_1,s_2} = log_2 \left[\frac{P(s_1, s_2)}{P(s_1) \,.\, P(s_2)} + 1 \right] \tag{8}$$

where $P(s_1, s_2)$ is the combine probability that s_1 and s_2 co-occur within a particular circumference, and, $P(s_1)$ and $P(s_2)$ are the respective probability of occurrence of terms s_1 and s_2. Such a measurement techniques account for relative positions of terms in a document. Further, in the sense of word order (such as "program executing" or "executing program"), asymmetric order is preferable, where $P(s_1, s_2)$ refers to the probability that s_2 exactly follow the s_2. The mutual information will be: zero if there is no co-occurrence, equal to one if terms s_1 and s_2 are distinct, and, equal to $log_2 \left(\frac{1}{P(s_1)} + 1 \right)$ if s_2 is absolutely correlated to s_2.

The drawback of the above formulation is that it can favor infrequent co-occurring terms as compared to frequent distant-occurring terms.

As another option, we can adopt the general description of conditional probability for calculating the stability of association between terms s_1 to s_2:

$$P(s_1|s_2) = \frac{P(s_1, s_2)}{P(s_2)} \tag{9}$$

This well known approach [22] is identical to the association rule used in data mining problem [6, 258]. Association rules have been used widely for identifying the expansion feature correlation with the user query terms [246,152].

Another corpus-based term similarity measure based on information content-based measurement is Resnik similarity [216]. Resnik measures the frequent information as information content (IC) of the Least common sub-sumer (LCS) (the closest common ancestor node of two synsets). The value of Resnik similarity would be greater than and equal to zero. The Resnik similarity can be formulated as

$$Sim_{resnik}(s_1, s_2) = -log\, p(LCS(s_1, s_2)) \tag{10}$$

The -ve sign makes the similarity score +ve because probabilities are always between [0,1].

Recently, Wikipedia has become popular for short query expansion. It is feasible in Wikipedia to have distinct articles with a common title. Every article describes the individual sense of the term, corresponding to the polysemous occurrences of the term in natural language. For example, term "apple" has two articles in Wikipedia, one indicating it as as a fruit and the other as the company. Reference [9] uses the Wikipedia for semantic enhancement of short query and measures the semantic similarity between two articles s_1 and s_2 as

$$Sim_{s_1,s_2} = \frac{|I(s_1) \cap I(s_2)| + |O(s_1) \cap O(s_2)|}{|I(s_1) \cup I(s_1)| + |O(s_2) \cup O(s_2)|} \tag{11}$$

where $I(s_1)$ & $I(s_2)$ is the set of articles that point to s_1 & s_2 respectively (as in-link) and $O(s_1)$ & $O(s_2)$ is the set of articles that s_1 & s_2 point to (as out-link).

Table 4 summarizes the mathematical form of term similarity score in One-to-One association based on above discussion.

Table 4: Summery of One-to-One Association for Term Ranking based on the term similarity score

Reference	Approaches	Mathematical form
Jaccard 1912 [124]	Jaccard coefficient	$\frac{df_{s_1 \wedge s_2}}{df_{s_1 \vee s_2}}$
Dice 1945 [85]	Dice coefficient	$\frac{2 \,.\, df_{s_1 \wedge s_2}}{df_{s_1} + df_{s_2}}$
Attar and Fraenkel 1977 [16]	Cosine similarity	$\frac{\sum_{d_j} w_{s_1,j} \cdot w_{s_2,j}}{\sqrt{\sum_{d_j} w^2_{s_1,s_1} \cdot \sum_{d_j} w^2_{s_2,s_2}}}$
Church and Hanks 1990 [64]	Mutual Information	$log_2 \left[\frac{P(s_1,s_2)}{P(s_1) \,.\, P(s_2)} + 1 \right]$
Wu and Palmer 1994 [274]	Wu & Palmer similarity	$\frac{2 \,.\, d(LCS)}{d(s_1) + d(s_2)}$
Resnik 1995 [216]	Resnik similarity	$-log\, p(LCS(s_1, s_2))$
ALMasri et al. 2013 [9]	Semantic similarity	$\frac{\lvert I(s_1) \cap I(s_2) \rvert + \lvert O(s_1) \cap O(s_2) \rvert}{\lvert I(s_1) \cup I(s_1) \rvert + \lvert O(s_2) \cup O(s_2) \rvert}$

3.2.2 One-to-Many Associations

In one-to-one association, each query term is expanded into correlated terms independently, whereas, in one-to-many association, multiple query terms can be expanded together – as a unit – into correlated terms. For example, consider queries "engineering technology" and "music technology". Now, word "technology" is frequently associated with word "information". Hence, an automatic expansion of "technology" in query "data technology" to "information technology" may work well because "information" is strongly correlated to the overall meaning of query "data technology". However, "information" does not relate to the overall meaning of "music technology". The main issue with one-to-one association is that it may not properly demonstrate the connectivity between the expansion term to the query as a whole. This issue has been discussed in reference [21], which deals with query-specific contexts instead of user-centric ones along with the context around and within the query. For resolving such language ambiguity problem, one-to-many association plays a crucial role.

References [114,115] use one-to-many association. Here, it is compulsory to correlate a new term extracted form the combination of ConceptNet and WordNet to a minimum of two original query terms before including the new term into expanded query. Let q be the original query and let s_2 be an expansion term. In one-to-many association, we may calculate the correlation coefficient of s_2 with q as:

$$\begin{aligned} c_{q,s_2} &= \frac{1}{|q|} \sum_{s_1 \in q} c_{s_1,s_2} \\ &= \frac{1}{|q|} \sum_{s_1 \in q} \sum_{d_j} w_{s_1,j} \,.\, w_{s_2,j} \end{aligned} \tag{12}$$

Other works based on one-to-many association are [214,277,75,22,252,217,32,99,149]. As a special mention, references [214,277] have gained large acceptance in literature because of their one-to-many association expansion feature and weighting scheme as described in Eq. 12.

Reference [214] uses Eq. 12 for finding pairwise correlations between terms in the entire collection of documents. Weight of a term s in document d_j, denoted $w_{s,j}$ (as in Eq. 12), is computed as the multiplication of term frequency (tf) of term s in document d_j and the inverse term frequency (itf) of the document d_j. The itf of document d_j is defined as $itf(d_j) = log(\frac{T}{|d_j|})$, where $|d_j|$ is

the number of distinct terms in document d_j and T indicates the number of terms in the entire collection. This approach is similar to the inverse document frequency (idf) used for document ranking.

Reference [277], uses concepts (group of contiguous nouns) instead of individual terms while expanding query. Concepts are chosen based on term co-occurrence with query terms. Concepts are picked from top retrieved documents, but they are determined on the basis of top passage (fixed size text window) rather than the whole document. Here, equation Eq. 12 is used for finding the term-concept correlations (instead of term-term correlations), where $w_{s_1,j}$ is the number of co-occurrence of query term s_1 in j^{th} passage and $w_{s_2,j}$ is the frequency of concept s_2 in j^{th} passage. Inverse term frequency of passages and the concepts contained in the passages – across the entire corpus – have been considered for calculating the perfect term-concept correlation score. A concept has a correlation factor with every query term. To obtain the correlation factor of the entire query, correlation factors of individual query terms are multiplied. This approach is known as local context analysis [277].

One-to-one association technique tends to be effective only for selecting expansion terms that are loosely correlated to any of the query terms. However, if correlation with the entire query or with multiple query words need to be considered, one-to-many association should be used. For example, consider the strong correlation of "information" with "technology". Here, "information" may be an expansion term for query "food technology" or "financial technology". This is not a desired expansion because "information technology" and "food technology" are unrelated. One way to overcome this problems is by adding *context words* to validate term-term associations. For example, in case of adding "information" as an expansion term for query "food technology", association of "food" and "information" should be considered strong only if these terms co-occur together sufficiently high number of times. Here, "food" is a context word added to evaluate term-term association of "information" and "technology" in the context of query "food technology". Such context words can be extracted from a corpus using term co-occurrence [20,21,264,130] or derived from logical significance of knowledge base [153,77,158,42].

Reference [260] – using WordNet data source for query expansion – found that expansion using term co-occurrence techniques are commonly not efficient because it doesn't assure a reliable word sense disambiguation. Although, this issue can be resolved by evaluating the correlation between WordNet senses associated with a query term and the senses associated with its neighboring query term. For example, consider query phrase "incandescent light". In WordNet, the definition of synset of incandescent contains word light. Thus, instead of the phrase "incandescent light", we can consider the synset of incandescent. Reference [169] uses this approach for word sense disambiguation (WSD).

Consider example query "tropical storm." In WordNet, the sense of the word "storm" determined through hyponym of the synset {violent stome, storm, tempest} is "hurricane", whose description having the word "tropical". As a result, the sense of storm is determined correctly. Reference [169] determines the correlation value of the terms in a phrase using:

$$C_{s_1,s_2,...,s_n} = \frac{P(phrase) - \prod_{s_i \in phrase} P(s_i)}{\prod_{s_i \in phrase} P(s_i)} \tag{13}$$

where $s_1, s_2, ..., s_n$ are the terms in a phrase, P(phrase) indicates the probability of documents containing the phrase, $p(s_i)$ is the probability of the individual term s_i in the phrase and $\prod_{s_i \in phrase} P(s_i)$ indicates the probability of document having all terms in a phrase.

Another approach for determining one-to-many association is based on the combination of various relationships between term pairs through a Markov chain framework [67]. Here, words having the highest probability of relevance in the stationary distribution of the term network are selected for query expansion. For every individual query term, a term network is built that consists of a pair of correlated terms corresponding to different types of relations (namely synonym, hyponym, co-occurrence). Reference [172] proposed a positional language model (PLM) that incorporates term proximity evidence in a model-based approach. Term proximity was computed directly based on proximity-based term propagation functions. Reference [247] proposed Proximity Probabilistic Model (PPM) that uses a position-dependent term count to compute the number of term occurrences and term counts propagated from neighboring terms. Recently, article [130] considered

term-based information and semantic information as two features of query terms and presented an efficient ad-hoc IR system using topic modeling. Here, first topic model is used for extracting the latent semantic information of the query term and then, term-based information is used as in a typical IR system. This approach is sturdier in relation to data paucity and it does well on large complicated (belonging to multiple topics) .

To overcome the limitations of considering term-to-term relationships – whether one-to-one or one-to-many – one can break the original query as one or more phrases, and then seek for phrases that are similar to it. Phrases usually offer richer context and have less ambiguity in comparison to their individual constituent words. At times, query expansion even at phrase level may not offer desired clarity, because the phrase may be compositional or non-compositional. With compositional phrases each and every term associated with the phrase can be expanded using similar alternative terms; the final expanded phrase keeps its significance. Reference [75] analyzes the phrases using n-grams from user's query logs. They filter the phrases that are not present in the documents being searched. Reference [171] selects most appropriate phrases for query expansion based on conceptual distance between two phrases (obtained using WordNet). First, phrases similar to the query phrase are selected as candidate phrases. Then, candidate phrases having low conceptual distance with respect to query phrase are considered in the set of most appropriate phrases. Recently reference [7] presented a query phrase expansion approach using semantic annotations in Wikipedia pages. It tries to enrich the user query with the phrases that disambiguate the original query word. However, generally it has been shown that short phrases have a more authentic representation of the information needed, e.g., "artificial intelligence". Further, phrases have a greater inverse document frequency in document collections in the corpus because when compared to individual query terms. This is because individual query terms occur more frequently in the document collection than the phrase as a whole. Reference [90] acknowledges that retrieval results are improved when pseudo relevance feedback is also included in query expansion based on phrases.

Dealing with idiomatic phrases can be troublesome. They are non-compositional in nature and replacing a word with a similar meaning word – as often done during query expansion – can completely change the meaning of the phrase. For example, "break a leg" is a theatrical slang meaning "good luck!". When we replace "leg" with synonym "foot", phrase "break a foot" gives an entirely different meaning from the original phrase.

Table 5 summarizes the mathematical form of term-term correlation value in one-to-many association based on the above discussion.

Table 5: Summary of one-to-many Association for term ranking based on the term-term correlation values

Publications	Approaches	Mathematical form
Qiu and Frei 1993 [214], Xu and Croft 1996 [277], Cui et al. 2003 [75], Bai et al. 2005 [22], Sun et al. 2006 [252], Riezler et al. 2008 [217], Bhatia et al. 2011 [32], Gan and Hong 2015 [99], Kuzi et al. 2016 [149]	Correlation coefficient	$c_{q,s_2} = \frac{1}{\lvert q\rvert}\sum_{s_1\in q} c_{s_1,s_2}$ $= \frac{1}{\lvert q\rvert}\sum_{s_1\in q}\sum_{d_j} w_{s_1,j}.w_{s_2,j}$
Liu et al. 2004 [169]	Correlation value	$\frac{P(phrase) - \prod_{s_i\in phrase} P(s_i)}{\prod_{s_i\in phrase} P(s_i)}$

3.2.3 Feature Distribution of Top Ranked Documents

Approaches discussed in this section are entirely distinct from the approaches described in earlier sections, because the query expansion techniques used in this section are not directly associated with the terms (individual or multiple) in the original query. This section uses the top relevant documents for query expansion in response to the initial query. The idea for using the top retrieved documents as a source of potential relevant documents for a user's motive comes from reference [16]. The top documents are retrieved in response to the initial query and have more detailed information about the initial query. This detailed information can be used for extracting the most relevant terms for expanding the initial query. Such query expansion approaches demonstrate

collectively better result in comparison to the above approaches. They can be subdivided into two categories:

- Query expansion through *Relevance feedback*. Query expansion terms are extracted from the retrieved documents in response to the initial query and the user decides the relevance of the results.
- Query expansion through *Pseudo-relevance feedback*. Query expansion terms are extracted from the top ranked documents in response to the initial query.

Relevance feedback (RF) is the most effective query expansion technique for modification of initial query using terms extracted from the documents in response to the initial query. The user is asked to assess the relevance of documents retrieved in response to the initial query. Retrieved documents are mostly shown to the user in some surrogate form, such as title, abstract, keywords, key-phrases or summaries. The user may also have a choice to see the entire documents before making their relevant judgment and selecting the relevant documents. After the user indicates relevant documents, these relevant documents are considered for extracting the terms for the initial query expansion. The top weighted terms are either added to the initial query automatically or based on manual selection by the user.

Quite a few term selection techniques have been proposed for query expansion, which are based on relevance feedback. The common thought behind the all similar techniques is to select terms that will describe the full meaning of the initial query. One of the first approaches, who investigated the relevance feedback is known as Rocchio's method [229]. This method used an information retrieval system based on the vector space model. The main idea behind this approach is to update the user's initial query vector based on the user's feedback. This method modifies the initial query vector as

$$\overrightarrow{q'} = \alpha.\overrightarrow{q} + \beta.\frac{1}{|RD|}\sum_{\overrightarrow{d_i}\in RD}\overrightarrow{d_i} - \gamma.\frac{1}{|ID|}\sum_{\overrightarrow{d_j}\in ID}\overrightarrow{d_j} \tag{14}$$

where:
$\overrightarrow{q'}$ is the modified query vector,
$\overrightarrow{q}$ is initial query vector,
α, β, γ manage the comparative importance associated with documents as initial Query Weight, relevant Documents (RD) Weight and irrelevant Documents (ID) Weight respectively, and
$\overrightarrow{d_i}, \overrightarrow{d_j}$ is relevant and irrelevant document vector respectively.
In the above paper (i.e., [229]) only the positive feedback documents and their terms were used to modify and expand the initial queries. Hence, the weights are typically set as $\alpha = 1, \beta = 0.75, \gamma = 0.15$ and, any negative term weights are neglected and set to 0.

Reference [132] presents a probabilistic model for calculating document matching score and comes up with superior results on TREC Programme collections. Here, first it retrieves the relevant documents in response to the user's initial query. Then, the documents matching score (MS) is computed as:

$$MS = \sum_{t_i \in q} \frac{tf_i \times (k_1 + 1)}{tf_i + NF \times k_1} \times w_i \tag{15}$$

where:
t_i is an individual term in the user's initial query q,
k_1 is the term frequency normalization factor,
tf_i is the term frequency of an individual term t_i in the document,
NF is document length normalization factor calculated as $NF = (1-c) + c \times \frac{DL}{AVDL}$ (c is a tuning constant, DL is the document length, and $AVDL$ is average document length)
w_i is the collection frequency weight of term t_i calculated as $w_i = log\frac{D_N}{n_i}$ (D_N is the total number of documents in the whole collection and n_i is the number of documents containing the term t_i).

Some of other works based on probabilistic reweighting formulation are [225, 227].

After the user selects relevant documents in response to the initial query, the system extracts all terms of these documents and ranks them according to Offer Weight (OW) computed as:

$$OW = r \times RW \tag{16}$$

where r is the number of relevant documents holding the query expansion terms and RW is the relevance weight.

RW is calculated as:

$$RW = log\frac{(r+0.5)(D_N - n - D_R + r + 0.5)}{(D_R - r + 0.5)(n - r + 0.5)} \tag{17}$$

where:
D_N is the total number of documents in the collection,
D_R is number of documents selected as relevant by the user, and
n is the number of documents containing the term.

After this, either the system asks the user to select relevant terms, or adds a fixed number of terms to the user's initial query (automatic query expansion).

An approach similar to relevance feedback approach is Pseudo-relevance feedback (or blind feedback, or retrieval feedback). This directly uses the top retrieved documents in response to user's initial query for composing query expansion terms. The user is not involved here in selection of relevant documents. Rocchio's method [229] can also be applied in the context of pseudo-relevance feedback, where every individual term extracted from the top retrieved documents is assigned a score by employing a weighting function to the entire collection. The score gathered by every individual term is estimated and the top terms are selected on the basis of resulting score. The Reccho's weights can be computed as

$$Score_{Rocchio}(t) = \sum_{d \in R} w(t,d) \tag{18}$$

where $w(t,d)$ indicate the weight of term t in pseudo-relevance document d and R is the set of pseudo-relevance documents.

However, a disadvantage of the above approach is that it considers the score of each term in document collection, in the process, showing more importance of the whole collection instead of the importance of the user's query. This problem can be resolved by analyzing the term distribution difference between the pseudo-relevant documents and the entire document collection. It is expected that terms having less information content will have nearly the same distribution in any documents in the whole collection. Terms that are closely related to the user's query will have a more probability of occurrence in the retrieve relevant documents.

Various term ranking functions have been proposed on the basis of term distribution in pseudo-relevant documents. These functions assign a high score to the terms that differentiate the relevant documents from the irrelevant ones. Some of the important term ranking functions have been described below.

Reference [225] proposes a weighting function known as Binary Independence Model (BIM) that assigns a score to the query terms for term ranking as follows:

$$Score_{BIM}(t) = log\frac{p(t|D_R)[1 - p(t|D_C)]}{p(t|D_C)[1 - p(t|D_R)]} \tag{19}$$

where $p(t|D_C)$ and $p(t|D_R)$ signify the probability of occurrence of the term t in the document collection D_C and a set of pseudo-relevant documents D_R respectively.

On the same lines, reference [86] uses a weighting function known as chi-square (χ^2) for scoring the query terms. It is formulated as:

$$Score_{\chi^2}(t) = log\frac{[p(t|D_R) - p(t|D_C)]^2}{p(t|D_C)} \tag{20}$$

Reference [224] presents a term selection method based on term weight known as Robertson selection value (RSV). It assigns a weight to a term on the basis of deviation in term distribution in the top retrieved documents. The term scoring method is formulated as:

$$Score_{RSV}(t) = \sum_{d \in R} w(t,d).\ [p(t|D_R) - p(t|D_C)] \tag{21}$$

where symbols have their meaning as given in equations 18 and 19.

On the same lines, reference [57] uses the Kullback-Leibler divergence (KLD) for measuring the term distribution difference between pseudo-relevant documents and the entire documents collection. Then, the terms having higher scores are added to the query along with KLD score as score of the term. The score of a term using KLD is computed as:

$$Score_{KLD}(t) = \sum_{t \in V} p(t|D_R).log\frac{p(t|D_R)}{p(t|D_C)} \tag{22}$$

where symbols have their meaning as given in equation 19.

Using the above term scoring approaches in query expansion, experimental studies [57,269,183] showed very good results.

In 2012, reference [95] presented a novel collaborative semantic proximity measurement technique known as PMING distance (further updated in [94]). It is based on the indexing information returned by search engines. It uses the number of occurrences of a term or a set of terms and counts the number of retrieved results returned by search engines.

The PMING distance is defined as the weighted combination of Pointwise Mutual Information (PMI) and Normalized Google Distance (NGD). Whereas PMI offers excellent performance in clustering, NGD gives better results in human perception and contexts. Overall, NGD and PMI exhibit good performance in capturing the semantic information for clustering, ranking and extracting meaningful relations among concepts. In order to understanding the PMING distance, we introduce some other concept similarity measurement technique such as PMI and NGD.

Pointwise Mutual Information (PMI) [64] is a point-to-point measure of association used in information theory and statistics. Actually, Mutual Information (MI) (Eq. 8) is a superset of PMI; PMI refers to individual event, while MI refers to the average of all possible events. Hence, PMI is defined as same as MI:

$$PMI_{s_1,s_2} = log_2\left[\frac{P(s_1,s_2)}{P(s_1)\,.\,P(s_2)}\right] \tag{23}$$

Normalized Google Distance (NGD) [65] measures the semantic relation between similar concepts that occur together in a number of documents retrieved by a query on Google or any other search engine. Originally, NGD was developed for Google, but it can be applied for any other search engine. NGD between the two terms s_1 and s_2 is defined as

$$NGD_{s_1,s_2} = \frac{max\{log\,f(s_1), log\,f(s_2)\} - log\,f(s_1,s_2)}{log\,N - min\{log\,f(s_1), log\,f(s_2)\}} \tag{24}$$

where:
$f(s_1)$, $f(s_2)$, and $f(s_1,s_2)$ denotes the number of results returned by search engine for query sets $\{s_1\}$, $\{s_2\}$ and $\{s_1,s_2\}$ respectively, and,
N is the total number of documents indexed by the search engine

N is usually unknown and varies very frequently. Hence, it can be approximated by a value significantly greater than $max\{f(s_1), f(s_2)\}$.

Though, in human perception NGD may stand good as a proximity measurement technique, in strict sense it cannot be considered as a metric because it does not satisfy the property of triangular inequality.

PMING distance [95,94] includes the combination of two semantic similarity measurement techniques: PMI and NGD. In PMING distance PMI & NGD are locally normalized and PMING distance is defined as a convex linear combination of the two locally normalized distances. While combining these two normalized distances, their relatives weights are chosen based on the context of evaluation using, e.g., Vector Space Model (VSM). For two terms s_1 and s_2 such that $f(s_1) \geq f(s_2)$, PMING distance between s_1 and s_2 in context W is given as a function $PMING : W \times W \rightarrow [0,1]$ and defined as:

$$PMING_{s_1,s_2} = \rho\left[1 - \left(log\frac{f(s_1,s_2)N}{f(s_1)f(s_2)}\right)\frac{1}{\mu_1}\right] + (1-\rho)\left[\frac{log\,f(s_1) - log\,f(s_1,s_2)}{(log\,N - log\,f(s_2))\mu_2}\right] \tag{25}$$

where:
ρ is a parameter to balance the weight of components such that $0 \leq \rho \leq 1$,
N is the total number (if known) or estimated number (if not known) of documents indexed by

the search engine,
μ_1 and μ_2 are constants; their value depends on the context of evaluation W and is defined as:

$$\mu_1 = \max_{s_1,s_2 \in W} PMI_{s_1,s_2}$$

$$\mu_1 = \max_{s_1,s_2 \in W} NGD_{s_1,s_2}$$

PMING offers the advantages of both PMI and NGD: it outperforms the state-of-the-art proximity measures in modeling human perception, modeling contexts and clustering of semantic associations – regardless of the search engine/repository.

Recently, article [201] presented a scoring function that uses two key properties of a query term: the number of feedback documents having the query term, and, the rarity of the query term in the whole document collection. The scoring function defined as:

$$Score(t, F^q) = log_2(df(t, F^q)) \times idf(t, C) \tag{26}$$

where:
F^q is the set of feedback document for the query q,
$df(t, F^q)$ indicates the number of documents in F^q having the term t
idf stand for inverse document frequency defined as $idf(t, C) = log \frac{N}{df(t,C)}$ (N is the number of documents in the whole collection and $df(t, C)$ indicates the document frequency of the term t in the collection C).

Every term ranking method has its own justification and the outcomes offered by their utilization are also distinct. In case of specific queries, it has been observed that organized sets of expansion terms recommended for each query are mostly unrelated to the original query [59]. However, various experimental analysis (such as [109,236,57,183]) observe that the selection of the ranking approach commonly does not have a huge significance on the system efficiency; it is just an approach to determine the set of terms for query expansion.

3.2.4 Query Language Modeling

In this approach for query expansion, a statistical language model is constructed that assigns a probability distribution over term-collections. Terms with maximum probability are chosen for query expansion. This approach is also known as model-based approach. The two popular foundation language models are relevance model (based on probabilities of terms in relevant documents) [156,70] and mixture model [286]; both utilize top retrieved documents for query expansion.

In relevance-based language model, reference [156] has caught the attention of researchers with its strong probabilistic approach. It assumes that the query q_i, and the top relevant documents set d are sampled randomly (identically and independently) from an unknown relevance model M_{rel}. It determines the probability of a term in relevant documents collection on the basis of its co-occurrence with the query terms. For approximating this relevance model, the probability of term t is computed using conditional probability of the initial query term q_i ($i \in 1, ..., n$). The probability of term t in the relevant documents is computed as:

$$p(t|M_{rel}) = \sum_{\theta_d \in R} p(\theta_d)\, p(t|\theta_d) \prod_{i=1}^{n} p(q_i|\theta_d) \tag{27}$$

In this equation, it has to be assumed that the term t and the query q_i are mutually independent once they elect a unigram distribution θ_d. Recently this relevance model has been used widely in query expansion. This model does not depend upon the distribution difference analysis, hence, it can be said that conceptually this model is very much like Rocchio method [229]. The main difference of this model from the Rocchio is that the top retrieved documents are assigned a weight such that the lower ranked documents have insignificant impact on term probability [155]. In the research area of relevance model, several studies have been published [173,30,183]. Reference [172] performed a correlative analysis on several states of pseudo relevance feedback and concluded that relevance model is the most efficient method for selection of expansion terms. Reference [30] uses external resources for generating features for weighing different types of query concepts and

considers the latent concepts for expanding the initial query. Reference [183] proposed a proximity-based feedback model that is based on the traditional Rocchio's model, known as PRoc. It focus on the proximity of terms rather than the positional information (unlike position relevance model (PRM)). It calculates the weights of candidate expansion terms by taking their distance from query terms into account. Reference [182] considers term dependencies during query expansion and the expansion technique is based on Markov random fields model. This model provides a powerful framework that includes both term occurrence and proximity-based features. An example of Markov random field is how many times the query terms occur within a window of fixed size in an organized or unorganized way. Reference [173] presents a technique for extracting the expansion terms from the feedback documents known as positional relevance model. Here, the focus is on query topics based on the positions of query terms in feedback documents. Another step in improving the research on relevance model [76] presents a neighborhood relevance model that uses the relevance feedback approaches for recognizing the specialty of entity linking across the document and query collections. Actually the main objective of entity linking is to map a string in a document to its entity in knowledge base and recognize the disambiguating context inside the knowledge base. Recently article [78] proposed a context dependent relevance model that provides an approach to incorporate the feedback through improvement of the document language models. For evaluating document language models, it uses the context information on the relevant or irrelevant document to obtain weight count (using BM25 weights [226, 223]) of the individual query terms.

Discussing mixture model method, reference [286] considers the top ranked documents extracted from the document collection that have both relevant and irrelevant information. The proposed method is a mixture productive model that integrates the query topic model $p(t|\theta_Q)$ to the collection language model $p(t|C)$. The collection language model is a suitable model for irrelevant information (content) in top-ranked documents. Following this mixture model, the log-likelihood for top-ranked documents is defined as

$$log\, p(T_R|\theta_Q) = \sum_{D \in T_R} \sum_{t} c(t, D)\, log(\lambda\, p(t|C) + (1 - \lambda)\, p(t, \theta_Q)) \tag{28}$$

where:
T_R is the set of top-ranked documents,
θ_Q is the estimated query model,
$c(t, D)$ is the number of occurrences of term t in document D, and
λ is a weighting parameter with a value between 0 and 1.

After the evaluation of log-likelihood, EM algorithm [82] is used to estimate the query topic model so that the likelihood of the top-ranked documents is maximized. However, estimating the query topic model is perhaps more difficult than estimating the document model because queries are generally short resulting in inadequacy of retrieved documents. Comparatively, this mixture model has a stronger theoretical justification, estimating the value of weighting parameter λ can be a difficult task.

Table 6 summarizes some important term similarity score in mathematical form for term ranking based on above discussion.

3.3 Selection of Query Expansion Terms

In the previous section 3.2, weighting and ranking of expansion terms have been done. After this step, the top-ranked terms are selected for query expansion. The term selection is done on individual basis; mutually dependence of terms is not considered. This may be debatable, however, some experimental studies (e.g, [165]) suggest that the independence assumption may be empirically equitable.

It may happen that chosen query expansion technique produces a large number of expansion terms, but it might not be realistic to use all of these expansion terms. Normally, only a limited number of expansion terms are selected for query expansion. This is because the information retrieval effectiveness of a query with a small set of expansion terms is usually better than the query having a large set of expansion terms [236]; this happens due to noise reduction.

Some experimental studies suggested the addition of optimum number of expansion terms in the initial query [109, 222, 52, 11, 61, 31, 269, 201, 290](described briefly in Section 3.1.1 and in

Table 6: Summery of Approaches for Term Ranking based on the term similarity score

Reference	Approaches	Mathematical form
Rocchio 1971 [229]	Recchio's weights	$\sum_{d \in R} w(t,d)$
Robertson and Jones 1976 [225]	Dice coefficient	$log \frac{p(t\|D_R)[1-p(t\|D_C)]}{p(t\|D_C)[1-p(t\|D_R)]}$
Doszkocs 1978 [86]	Chi-square (χ^2)	$log \frac{[p(t\|D_R)-p(t\|D_C)]^2}{p(t\|D_C)}$
Robertson 1990 [224]	Robertson selection value (RSV)	$\sum_{d \in R} w(t,d).\ [p(t\|D_R) - p(t\|D_C)]$
Carpineto et al. 2001 [57]	Kullback-Leibler divergence (KLD)	$p(t\|D_R).\ log \frac{p(t\|D_R)}{p(t\|D_C)}$
Zhai and Lafferty 2001 [286]	Log-likelihood	$log\, p(T_R\|\theta_Q) = \sum_{D \in T_R} \sum_t c(t,D)\, log(\lambda\, p(t\|C) + (1-\lambda)\, p(t,\theta_Q))$
Cilibrasi and Vitanyi 2007 [65]	Normalized Google Distance (NGD)	$NGD_{s_1,s_2} = \frac{max\{log\, f(s_1), log\, f(s_2)\} - log\, f(s_1,s_2)}{log\, N - min\{log\, f(s_1), log\, f(s_2)\}}$
Franzoni and Milani 2012 [95], Franzoni 2017 [94]	PMING distance	$PMING_{s_1,s_2} = \rho \left[1 - \left(log \frac{f(s_1,s_2)N}{f(s_1)f(s_2)}\right) \frac{1}{\mu_1}\right] + (1-\rho)\left[\frac{log\, f(s_1) - log\, f(s_1,s_2)}{(log\, N - log\, f(s_2))\mu_2}\right]$
Paik et al. 2014 [201]	Scoring function	$Score(t, F^q) = log_2(df(t, F^q)) \times idf(t, C)$

Table 7). However, this suggested optimum can vary from a few terms to a few hundred terms. The expansion terms of the relevant or top-ranked documents improve the effectiveness of query expansion by 7% to 25% [52]. On the contrary, some studies shows that the number of terms used for query expansion is less important than the terms selected on the basis of types and quality [241]. It has been commonly shown that the effectiveness of the query expansion decreases minutely with the non-optimum number of expansion terms [57]. Most of the experimental studies observed that the number of expansion terms is of low relevance and it varies from query to query[39]. It has been observed that the effectiveness of query expansion (measured as mean average precision) decreases when we consider less than 20 expansion terms [201,290]. Usually 20-40 terms is the best choice for query expansion. Reference [286] assigns a probability score to each expansion term and selects those with score greater than a fixed threshold value p=0.001.

Table 7: Summery of Terms selection suggested by several researchers

Number of Terms	Reference
One third of the terms	Robertson and Willett 1993 [222]
5 to 10 terms	Amati et al. 2003 [11], Chang et al. 2006 [61]
20 to 30 terms	Harman 1992 [109], Zhang et al. 2016 [290]
30 to 40 terms	Paik et al. 2014 [201]
Few hundreds terms	Bernardini and Carpineto 2008 [31], Wong et al. 2008 [269]
350 to 530 terms	Buckley et al. 1995 [52]

However, instead of considering an optimum number of expansion terms, it may be better to adopt more aware selection techniques. Several experimental results notice that the optimum number of expansion terms vary from query to query [50,38,55]. Focus on selection of most relevant terms for query expansion instead of optimum number of terms yields better results [59,55].

For the selection of the expansion terms on the basis of ranks assigned to the individual term, various approaches have been proposed that exploit additional information. Reference [59] proposed a technique that uses multiple term ranking functions and selects the most common terms for each query. A similar approach is utilized in [68], however, multiple feedback models are constructed from the same term ranking function. This is done by reconsidering documents from the corpus and by creating alternatives of the initial query. The paper also claims that the proposed technique is effective for eliminating the noise from expansion terms. It aims to expand the query terms that are related to the various query features. Another approach for selecting expansion terms that depends upon the query ambiguity has been proposed in reference [62]. Here, the number of expansion terms depend on the ambiguity of the initial query in the web or the user log; the ambiguity is determined by the clarity score [72]. Reference [55] uses a classifier to recognize the relevant or irrelevant expansion terms. Whether the classifier parameter works well or not for labeling the individual expansion terms, depends on the effectiveness of retrieval performance and

co-occurrence of query terms. Their study shows that top retrieved documents contain as many as 65% harmful terms. For selecting the best expansion terms Reference [66] optimized the retrieved data with respect to uncertainty sets resulting in an optimization problem.

However, it has been shown that majority of existing works on QE [155,270] only focused on indexing and document optimization for selection of expansion terms and neglects the re-ranking score. However, recently a number of articles [83,290] supported the re-ranking with valid proof and obtained good retrieval effectiveness. Reference [270] proposed impact-sorted indexing technique that utilizes a special index data structure; the technique improves the scoring methods in information retrieval. Reference [155] uses the pre-calculated pairwise document similarities to reduce the searching time for expanded queries. However, supporting re-ranking, reference [83] points out that re-ranking can provide nearly identical performance as the results returned from the second retrieval done on expanded query. This works specifically for precision-oriented metrics. This has also verified in experimental result of reference [290], which utilizes re-ranking as the default approach for information retrieval.

3.4 Query Reformulation

This is the last step of query expansion, where the expanded query is reformulated to achieve better results when used for retrieving relevant documents. The reformulation is done based on weights assigned to the individual terms of the expanded query – known as query reweighting.

A famous query reweighting method was proposed in reference [235], which is influence by Rocchio's method [229] for relevance feedback and its consequential developments. It can be formulated as:

$$w'_{t,q_e} = (1-\lambda)\,.\,w_{t,q} + \lambda\,.\,W_t \tag{29}$$

where w'_{t,q_e} is the reweighting of term t of the expanded query q_e,
W_t is a weight assigned to the expansion term t, and,
λ is weighting parameter, which weight the comparative contribution of the original query (q) terms and the expansion terms.

When Rocchio's weights (see equation 18) are used for calculating the weights of the query expansion terms, that are extracted from the pseudo-relevant documents, it can be observed that the expanded query vector measured by equation 29 is relevant to the pseudo relevant documents. This reduces the term distribution difference between pseudo relevance documents and documents having expansion terms when terms are reweighed by Rochhio's weighing scheme. The intention is to assign low weight to a top ranked term (in an expanded query) if its relevance score with respect to the whole collection of documents is low. A number of experimental results support this observation for various languages [290,201,269,57], Hindi [33,201], Asian [237] and European languages [80,151,11]. It has been observed that the reweighting system based on inverse term ranks also provides a favorable outcome [116,59]. Another observation is that the document-based weights used for the original unexpanded query and the term distribution difference-based scores used for expansion terms have different units of measurement. Hence, before using them in equation 29 their values must be normalized. A number of normalization approaches have been discussed in survey [269] and it was observed that the discussed approaches commonly provide similar outcomes. However, reference [191] observes the need for a better approach that not only normalizes data but also increases equality among normalized terms, which can be more expressive.

In addition to the above discussion, the value of weighting parameter (λ) in equation 29 should be adjusted appropriately for improving retrieval effectiveness. A common choice is to grant more significance – double – to the user's initial query terms in comparison to the expanded query terms. Another way is to use the query reweighting formula without weighting parameter (λ) as suggested in reference [11]. Another effective approach is to query-wise compute weight to be assigned to the expansion terms. For example, references [172,173], use relevance feedback in combination with a learning approach to forecast the values of weighting parameter (λ) for each query and every collection of feedback documents. They also discuss various techniques – based on e.g., length, clarity and entropy – to measure the correlation of query terms with the entire collection of documents as well as with only feedback documents. However, equation 29 can also be used for extracting expansion terms from hand-built knowledge resources (such as thesaurus,

WordNet and ConceptNet). The weighting score may be assigned on the basis of attributes such as path length, number of co-occurrences, number of connections and relationship types [133]. For example, reference [260], uses expanded query vector with eleven concept types sub vectors. Each concept type sub vector that comes inside the noun part of WordNet is assigned individual weights. Examples of used concept type sub vectors are "original query terms" and "synonyms". Similarly, reference [115] uses activation score for weighting of expansion terms.

When document ranking is based on language modeling approach (see the section 3.2.4), the query reweighting step usually favorably expands the original query. In language modeling platform, most relevant documents are the ones that decrease Kullback-Leibler divergence (KLD) between document language model and the query language model. It is formulated as:

$$Sim_{KLD}(Q, D) \propto \sum_{t \in V} p(t|\theta_Q).\, log \frac{p(t|\theta_Q)}{p(t|\theta_D)} \tag{30}$$

where:
θ_Q is the query model (usually calculated using the original query terms), and,
θ_D is the document model.

Document model θ_D is calculated based on unknown terms via probability smoothing techniques, such as Jelinek-Mercer interpolation[128,129]:

$$p(t|\theta'_D) = (1 - \lambda)\,.\,p(t|\theta_D) \;+\; \lambda\,.\,p(t|\theta_C) \tag{31}$$

where:
$p(t|\theta'_D)$ is the probability of term t in θ'_D (documents retrieved using expanded query), and
θ_C is the collection model.

Equation 30 raises the following question: is it possible to build a better query model by obtaining similar terms with their concern probabilities? Further, will it smooth the original query model using the equivalent expanded query model (EQM) just as collection model θ_C smooths the document model based on Eq. 31 . To answer this, several approaches have been proposed to build an expanded query model that not only consider feedback documents [286,156], but also term relations [22,54,99], domain hierarchies [21] and can be heuristic [238]. Hence, reference [58] suggested that instead of considering a particular method, one can come up with a superior expanded query model (calculated using Jelinek-Mercer interpolation [129]) given as:

$$p(t|\theta'_Q) = (1 - \lambda)\,.\,p(t|\theta_Q) \;+\; \lambda\,.\,p(t|\theta_{EQM}) \tag{32}$$

where, for each term $t \in \theta'_Q$: $p(t|\theta'_Q)$ is the probability of term t in expanded query θ'_Q,
$p(t|\theta_Q)$ is the probability of term t in original query Q,
$p(t|\theta_{EQM})$ is the probability of term t in expanded query model θ_{EQM}, and,
λ is the interpolation coefficient.

This equation is the probabilistic representation of Eq. 29 and many articles [279,142,58,99,149,290] have used it for probabilistic query reweighting.

Though the query reweighting approach is generally used in query expansion techniques, it is not mandatory. For example, one can increase the number of similar terms that characterize the original query without using the query reweighting techniques [58]. Another way can be to first increase the number of similar query terms and then apply a customized weighting function for ranking the expanded query terms, in stead of using the fundamental weighting function used for reweighting the expanded query. This technique was used in reference [227] to enhance the Okapi BM25 ranking function [228]. Some other approaches for query reformulation are utilization of structured query [67,213,139,125], Boolean query [207,169,105,141], XML query [136,63,135] and phrase matching [14].

4 Classification of Query Expansion Approaches

On the basis of data sources used in query expansion, several approaches have been proposed. All these approaches can be classified into two main groups: (1) Global analysis and (2) Local analysis. Global and Local analysis can be further split into four and three subclasses respectively as shown in Fig. 5. This section discusses properties of various data sources used in query expansion as shown in Fig. 5; the query expansion approaches have been categorized based on these properties.

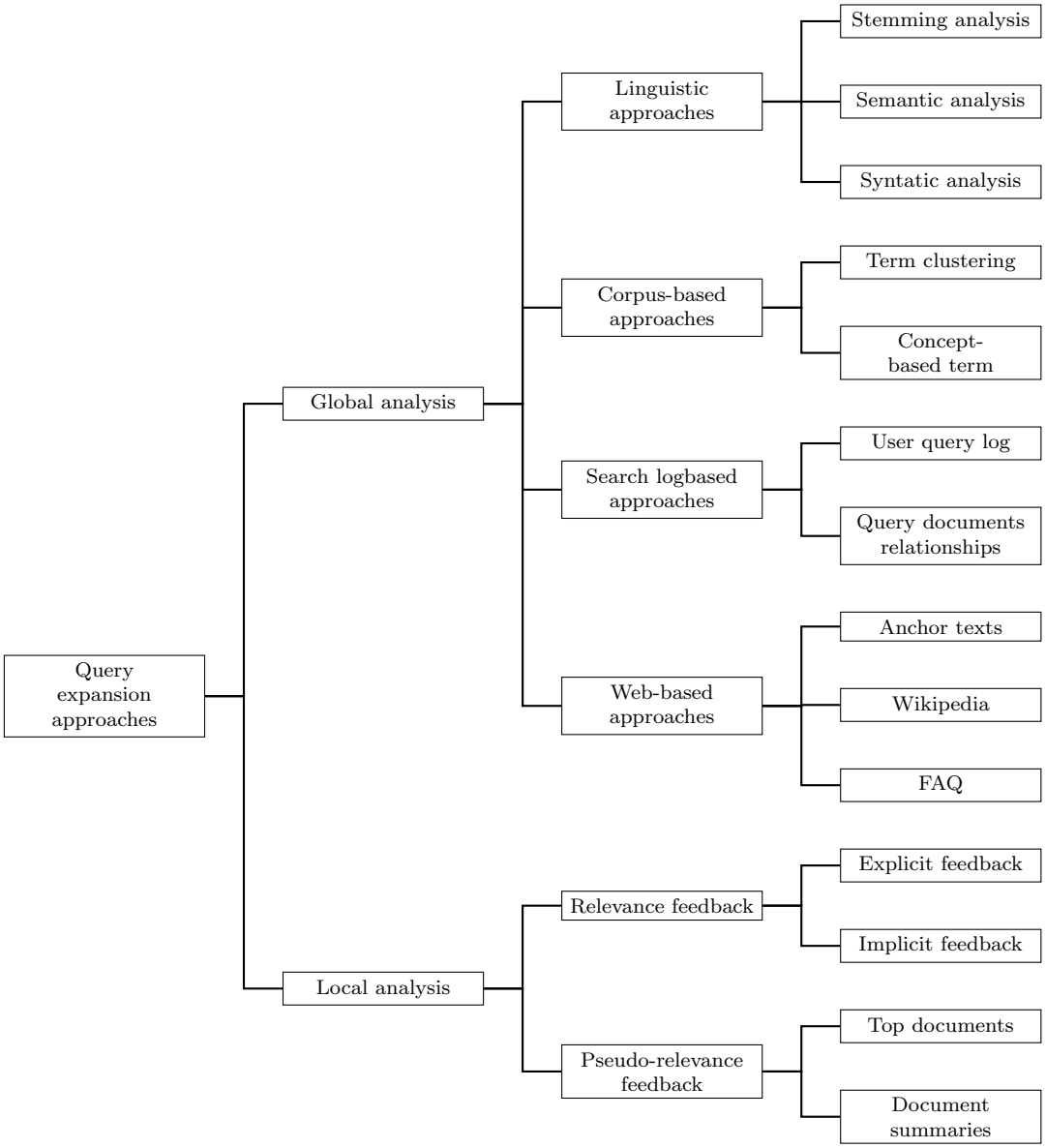

Fig. 5: Classification of query expansion approaches based on data sources.

4.1 Global Analysis

In global analysis, query expansion techniques implicitly select expansion terms from hand-built knowledge resources or from large corpora for expanding/reformulating the initial query. Only individual query terms are considered for expanding the initial query and expansion terms are semantically similar to the original terms. Each term is assigned a weight and expansion terms can be assigned less weight in comparison to the original query terms. Global analysis can be classified into four categories on the basis of query terms and data sources: linguistic-based, corpus-based, search log-based and web-based. Each approach has been discussed briefly in the following sections.

4.1.1 Linguistic-based Approaches

The approaches in this category analyze the expansion features such as lexical, morphological, semantic and syntactic term relationships to reformulate or expand the initial query terms. They use thesaurus, dictionaries, ontologies, Linked Open Data(LOD) cloud or other similar knowledge resources such as WordNet or ConceptNet.

Word stemming is one of the first and among the most influential query expansion approaches in linguistic association to reduce the inflected word from its root word. The stemming algorithm (e.g., [211]) can be utilized either at retrieval time or at indexing time. When used during retrieval, terms from initially retrieved document are picked, and then, these terms are harmonized to the morphological types of query terms (e.g., [145,200]). When used during indexing time, document word stems are picked, and then, these words are harmonized to the query root word stems (e.g., [122]). Morphological approach is an ordered way of studying internal structure of the word. It has been shown to give better result than the stemming approach [40,192], however, it requires querying to be done in a structured way.

Use of semantic and contextual analysis are other popular query expansion approaches in linguistic association. It includes knowledge sources such as Ontologies, LOD cloud, dictionaries and thesaurus. In the context of ontological based query expansion, reference [34] uses domain-specific and domain-independent ontologies. Reference [272] utilizes the rich semantics of domain ontology and evaluates the trade off between the improvement in retrieval effectiveness and the computation cost. Several research works have been done on query expansion using a thesaurus. WordNet is a well known thesaurus for expanding the initial query using the word synsets. As discussed earlier, many of the research works use WordNet for expanding the initial query. For example, reference [260] uses WordNet to find the synonyms. Reference [245] uses WordNet and POS tagger for expanding the initial query. However, this approach has some practical problems, such as no 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 query expansion is beneficial only if the query words are unambiguous in nature [104,260]; word sense disambiguation (WSD) is not an easy problem [195,203]. Several research works have attempted to address the WSD problem. For example, reference [196] suggests that instead of considering the replacement of the initial query term with its synonyms, hyponyms, and hyperonyms, it is better to extract the similar concept from the same domain of the given query from WordNet (such as the common nodes and glossy terms). Reference [103] uses the semantically similar information from WordNet in different group; this may be combined to expand the initial query. References [288,246, 169] combine WordNet concepts – that are extracted by consecutive application of heuristic rules to match the similar query terms – with other term extraction techniques. Reference [239] uses linguistic and semantic features of the initial query over linked data for QE as discuss in earlier section 2.2.2. Recently, reference [5] introduced a WSD algorithm based on random walks over large Lexical Knowledge Bases (LKB). The experiments give better results than other graph-based approaches when executed on a graph built from WordNet and extended WordNet. Nowadays, Word Embeddings techniques are being used widely for query expansion, e.g., as in references [230,84,149] discussed earlier.

Another important approach that improves the linguistic information of the initial query is syntactic analysis [289]. Syntactic based query expansion uses the enhanced relational features of the query terms for expanding the initial query. It expands the query mostly through statistical approaches [272]. It recognizes the term dependence statistically [218] by employing techniques such as term co-occurrence. Reference [252] 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 query expansion in a density and relation based passage retrieval framework. 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 [289] or integrate the taxonomic information effectively [171].

4.1.2 Corpus-based Approaches

Corpus-based Approaches examine the contents of the whole text corpus to recognize the expansion features utilized for query expansion. They are one of the earliest statistical approaches for query expansion. 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 expanded query. Corpus-based approaches have two admissible strategies: (1) term clustering [131,188,73], which groups document terms into clusters based on their co-occurrences, and, (2) concept based terms [214,

93,194], where expansion terms are based on the concept of query rather than the original query terms. Recently reference [149] selected the expansion terms after the analysis of the corpus using word embedding, where each term in the corpus is characterized with a vector embedded in a vector space.[290] uses four corpora as data sources (including one industry and three academic corpus) and present a Two-stage Feature Selection framework (TFS) for query expansion known as Supervised Query Expansion (SQE), discuss in earlier section 3.1.1 . Some of the other approaches established an association thesaurus based on the whole corpus; e.g., reference [101] uses context vectors, reference [57] uses the term co-occurrence, reference [116] uses the mutual information and reference [187] uses the interlinked Wikipedia articles.

4.1.3 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 based on the analysis of search logs. With the fast growing size of the web and increasing use of web search engines, the abundance of search logs and their ease of use have made them an important source for query expansion. It usually contains user queries corresponding to the URLs of Web pages. Reference [74] 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 [75] uses search logs for query expansion; their experiments give better results when compared with query expansion based on pseudo relevance feedback (PRF). 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 [267] that implicit measurements are relatively good, but, their performance may not be the same for all types of users and search task.

There are commonly two types of query expansion approaches used on the basis of web search logs. The first approach considers queries as documents and extracts features of these queries that are related to the user's initial query [118]. Among the techniques based on first approach, some use their combined retrieval results [119], while some do not (e.g., [118,282]). In the second approach, the features are extracted on relational behavior of queries. For example, reference [19] represents queries in a graph based vector space model (query-click bipartite graph) and analyzes the graph constructed by query logs. References [75,218,56] extract the expansion terms directly from clicked results. References [92,263] use the top results from past query terms entered by users. Under the second approach queries are also extracted from related documents [38,264], or through user clicks [280,282,117]. The second is more popular and has been shown to give better results.

4.1.4 Web-based Approaches

These approaches include anchor texts and Wikipedia for expanding the user's original query, and have become popular in recent times. Anchor text was first used in reference [180] for associating hyper-links with linked pages, as well as with the pages in which anchor texts are found. In the context of a web-page, 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 [144] used anchor texts for QE; their experimental results suggest that anchor texts can used to improve the traditional QE based on query logs. On similar lines, reference [79] suggested that anchor text can be an effective substitute for query logs. It demonstrated effectiveness of query expansion techniques using log-based stemming through experiments on standard TREC collection dataset.

Another popular approach is use of Wikipedia articles, titles and hyper-links (in-link and out-link) [14,9]. As we know, Wikipedia is the largest encyclopedia freely available on the web; articles are regularly updated and new ones are added every day. These features make it an ideal knowledge source for query expansion. Recently, quite a few research works have used it for query expansion (e.g., [163,14,279,3,9]). Article [7] attempts to enrich initial queries using semantic annotations in Wikipedia articles combined with phrases disambiguation. Experimental results show better results in comparison to the relevance based language model.

FAQs are another important web-based source of information for improving the QE. Recently published article [138] uses domain specific FAQs data for manual query expansion. Some other works using FAQs are [4,248,218].

4.2 Local Analysis

Local analysis includes query expansion techniques that select expansion terms from documents collection retrieved in response to the user's initial (unmodified) query. The working belief is that document retrieved in response to the user's initial query are relevant, hence, terms present in these documents should also be relevant to the initial query. Using local analysis, there are two ways to expand user's original query: (1) Relevance feedback and (2) Pseudo-relevance feedback. These two ways are discussed next.

4.2.1 Relevance Feedback (RF)

In this approach, user's feedback about documents retrieved in response the initial query is collected; the feedback is about whether or not retrieved documents are relevant to the user's query. The query is reformulated based on documents found relevant as per user's feedback. Rocchio's method [229] was amongst the first to use relevance feedback. Relevance feedback can further be categorized into two types: explicit feedback and implicit feedback. In explicit feedback, user explicitly evaluates the relevance of retrieved documents (as done in [235,109]), whereas in implicit feedback, user's activity on the set of documents retrieved in response to initial query is used to indirectly infer user preferences (e.g. as done in [62,293,100]). Relevance feedback suffers from lack of semantics in the corpus [272]. This restrains its applications in several occasions, for example, when the query concept is as general as a disjunction of more specific concepts (see Chap. 9 in [177]). Some of the research works based on relevance feedback are [51,236,231,177]; these have been discussed earlier in Sec. 3.2.3.

4.2.2 Pseudo-relevance Feedback (PRF)

Here, neither explicit nor implicit feedback of user is collected. Instead, the feedback collection process is automated by directly using the top ranked documents (or their snippets), retrieved in response to the initial query, for query expansion. Pseudo-relevance feedback is also known as blind feedback, or, retrieval feedback. It has been discussed briefly earlier in Sec. 3.2.3. This technique was first proposed in reference [71], which employs this technique in a probabilistic model. Reference [278] proposed "local context analysis" technique to extract the query expansion terms from the top documents retrieved in response to the initial query. Each of the candidate expansion term is assigned a score on the basis of co-occurrence of query terms. The candidate terms with highest score are selected for query reformulation. A recent work [243] uses fuzzy logic-based query expansion techniques and selects top-retrieved documents based on pseudo-relevance feedback. Here, each expansion term is assigned a distinct relevance score using fuzzy rules. Terms having highest scores are selected for query expansion. The experimental results demonstrate that the proposed approach achieves significant improvement over individual expansion, expansion on the basis of entire query and other related advanced methods. Reference [10] proposed deep learning based QE technique and compared it with PRF and other expansion models; the results show a notable improvement over other techniques using various language models for evaluation.

However, considering top retrieved may not always be the best strategy. For example, for a particular query, if the top retrieved documents have very similar contents, the expanded terms – selected from the top retrieved documents – will be also very similar. Hence, the expanded will not be useful for effective query expansion. Apart from using the top-ranked documents or their snippets, several other approaches have been proposed. For example, techniques based on passage extraction [277], text summarization [150], and document summaries [61]. Some of the other articles using PRF are [55,279,173]; these have been discussed in earlier sections.

Table 8 summarizes influential query expansion approaches in chronological order on the basis of five prominent features: Data Sources, Term Extraction Methodology, Term representation, Term Selection Methodology, and Weighting Schema.

Table 8: Summary of Research in the area of Query Expansion

Reference	Data Sources	Term Extraction Methodology	Term representation	Term Selection Methodology	Weighting Schema
Robertson 1990 [224]	Corpus	All terms in corpus	Individual terms	Swets model	match function
Qiu and Frei 1993 [214]	Corpus	All terms in corpus	Individual terms	Term-Concept similarity	Correlation based weights
Voorhees 1994 [260]	WordNet	Synsets & hyponyms of the query	Individual terms	Hyponym chain length	Vectors multiplication of query and concepts
Xu and Croft 1996 [277]	Corpus & top-ranked documents	Contiguous nouns in top retrieved passages	Phrases	Term co-occurrence	Ranked-based weights
Robertson et al. 1999 [228]	Top-ranked documents	All terms in top retrieve documents	Individual terms	Robertson selection value (RSV)	Probabilistic reweighting
Carpineto et al. 2001 [57]	Corpus & top-ranked documents	All terms in top retrieve documents	Individual terms	Kullback-Leibler divergence (KLD)	Rocchio & KLD scores
Zhai and Lafferty 2001 [286]	Corpus & top-ranked documents	All terms in top retrieve documents	Individual terms	Mixture model	Query language model
Lavrenko and Croft 2001 [156]	Corpus & top-ranked documents	All terms in top retrieve documents	Individual terms	Relevance model	Query language model
Cui et al. 2003 [75]	User logs & corpus	Query-documents correlation	Individual terms	Probabilistic term-term association	Cohesion weights
Billerbeck et al. 2003 [38]	Query logs	Query association	Individual terms	Robertson selection value (RSV)	Probabilistic reweighting
Kraft and Zien 2004 [144]	Anchor texts	Adjacent terms in anchor text	Phrases	Median rank aggregation	Unweighted terms
Liu et al. 2004 [169]	Corpus, top-ranked documents & WordNet	Phrase classification & WordNet concepts	Individual terms & phrases	Term co-occurrence & WSD	Boolean query
Bai et al. 2005 [22]	Top-ranked documents	Adjacent terms in top-ranked documents	Individual terms	Term co-occurrence & Information Flow (IF)	Query language model
Collins-Thompson andCall an 2005 [67]	WordNet, corpus, stemmer and top-ranked documents	Probabilistic term association network	Individual terms	Markov chain	Structured query
Sun et al. 2006 [252]	Corpus	Relevant contextual terms	Phrases	DBS & RBS	Correlation-based weights
Hsu et al. 2006 [114]	ConceptNet & WordNet	Terms having the same concept	Individual terms	Using discrimination ability & concept diversity	Correlation-based weights
Riezler et al. 2007 [218]	FAQ data	Phrases in FAQ answers	Phrases	SMT techniques	Unweighted terms
Metzler and Croft 2007 [182]	Corpus & top-ranked documents	Markov random fields model	Individual terms	Maximum likelihood	Expanded query graph
Bai et al. 2007 [21]	Corpus, user domains & top-ranked documents	Terms & nearby terms	Individual terms	Query classification & mutual information	Query language model

Table 9: Summary of Research in the area of Query Expansion (Cont. from Table 8)

Reference	Data Sources	Term Extraction Methodology	Term representation	Term Selection Methodology	Weighting Schema
Lee et al. 2008 [157]	Corpus & top-ranked documents	Clustering of top-ranked documents	Individual terms	Relevance model	Query language model
Cao et al. 2008 [55]	Corpus & top-ranked documents	All terms in top retrieve documents	Individual terms	Term classification	Query language model
Arguello et al. 2008 [14]	Wikipedia	Anchor texts in top retrieve Wikipedia documents	Phrases	Document rank & link frequency	Sum of entry likelihoods
Xu et al. 2009 [279]	Wikipedia	All terms in top retrieve articles	Individual terms	Relevance model	Query language model
Yin et al. 2009 [282]	Query logs & Snippets	All terms in top retrieve snippets & random walk on query-URL graph	Individual terms	Relevance model & mixture model	Query language model
Dang and Croft 2010 [79]	Anchor texts	Adjacent terms in anchor text	Individual terms & Phrase	Kullback-Leibler divergence (KLD)	Substitution probability
Lv and Zhai 2010 [173]	Corpus & top-ranked documents	All terms in top retrieve feedback documents	Individual terms	Positional relevance model (PRM)	Probabilistic reweighting
Kim et al. 2011 [141]	Corpus	All terms in top retrieve documents	Individual terms	Decision tree-based	Boolean query
Bhatia et al. 2011 [32]	Corpus	All terms in the corpus	Individual terms & Phrases	Document-centric approach	Correlation based weights
Miao et al. 2012 [183]	Corpus	All Proximity-based terms in the corpus	Individual terms	Proximity-based feedback model (PRoc)	KLD scores
Zhou et al. 2012 [293]	User logs	Associated terms extracted from top ranked documents	Individual terms	Annotations and resources the user has bookmarked	Correlation based weights
Aggarwal and Buitelaar 2012 [3]	Wikipedia & DBpedia.	Top ranked articles from best selected articles as concept candidates	Individual terms & phrases	Explicit Semantic Analysis (ESA) score	tf-idf
ALMasri et al. 2013 [9]	Wikipedia	In-link & out-link articles in top retrieve articles	Individual terms	Semantic similarity	Semantic similarity score
Pal et al. 2013 [205]	Corpus	All terms in top retrieve documents	Individual terms	Association & distribution based term selection	KLD score
Bouchoucha et al. 2013 [48]	ConceptNet	Diverse expansion terms from retrieve documents	Individual terms	MMRE (Maximal Marginal Relevance -based Expansion)	MMRE score
Augenstein et al. 2013 [17]	LOD cloud	Neighbors term in whole graph of LOD	Individual terms	Mapping Keywords	tf-idf

Table 10: Summary of Research in the area of Query Expansion (Cont. from Table 9)

Reference	Data Sources	Term Extraction Methodology	Term representation	Term Selection Methodology	Weighting Schema
Pal et al. 2014 [206]	WordNet	Synonyms, holonyms & meronyms of the query terms	Individual terms	pseudo relevant documents	Normalized weights
Paik et al. 2014 [201]	Top retrieved set of documents	All terms in feedback documents	Individual terms	Incremental Blind Feedback (IBF)	Scoring function
Al-Shboul and Myaeng 2014 [7]	Wikipedia	Titles & in/out links of Wikipedia pages	Individual terms & Phrases	Wikipedia page surrogates	Phrase likelihood model
Dalton et al. 2014 [77]	Wikipedia & Freebase	All terms in top ranked articles links to knowledge bases	Individual terms & phrases	Entity query feature expansion (EQFE)	Document retrieval probability
Anand and Kotov 2015 [13]	DBpedia & ConceptNet	Top-k terms in term association graphs	Individual terms	Information theoretic measures based on co-occurrence & mixture model	KLD score
Xiong and Callan 2015 [276]	Freebase	All terms in top retrieve documents	Individual terms	tf.idf based PRF & Freebase's entity categories	tf.idf & Jensen-Shannon divergence score
Gan and Hon 2015 [99]	Wikipedia & corpus	Top-k correlated terms in the corpus	Individual terms	Markov network model	Probabilistic query reweighting
Roy et al. 2016 [230]	Corpus	All nearby terms at word embedding framework	Individual terms	K-nearest neighbor approach (KNN)	Query language model
Diaz et al. 2016 [84]	Corpus	All terms in locally-trained word embeddings	Individual terms	local embedding approach	Query language models & KLD score
Singh and Sharan 2016 [243]	Top-ranked documents	All the unique terms of top N retrieved documents	Individual terms	fuzzy-logic-based approach	CHI score, Co-occurrence score, KLD score & RSV score
Zhang et al. 2016 [290]	Corpora	Top-k terms in retrieval documents	Individual terms	Two-stage Feature Selection (TFS)	Probabilistic query reweighting
Zhou et al. 2017 [295]	User profiles & folksonomy data	All semantic similar terms inside the enriched user profiles	Individual terms	Word embeddings & topic models	Topical weighting scheme

5 DISCUSSION AND CONCLUSIONS

This article has presented a comprehensive survey highlighting current progress, emerging research directions, potential new research areas, working methodology, and detail classification techniques used for query expansion. Although there is no perfect solution for the vocabulary mismatch problem in information retrieval system, query expansion has the capability to overcome the primary limitations. That is, provide the supporting explanation of the information needed for efficient information retrieval, which could not be provided earlier due to the unwillingness or inability of the user. As we see in the present scenario of the search systems, most frequent queries are still one, two or three words; the same as in the past few decades.

The lack of query terms increases the ambiguity in choosing among the many possible synonymous meanings of the query terms. This heightens the problem of vocabulary mismatch. This, in turn, has motivated the necessity and opportunity to provide intelligent solutions to vocabulary mismatch problem. Over the past few decades, a lots of research have been done in the area of query expansion based on data sources used, applications and expansion techniques. This article classifies the various data sources into three categories: documents used in the retrieval process, hand-built knowledge resources and external text collections and resources. Recently, it has been shown that data sources of type external text collections and resources are used widely for query expansion, specially, web data. In research involving web data, Wikipedia is an popular data source because it is the freely available largest encyclopedia on the web, where articles are regularly updated/added.

Expansion approaches can be manual, automatic or interactive, (such as linguistic, corpus-based, web-based, search log-based, RF and PRF); they expand the user's original query on the basis of query features and available data sources. Query characteristic depends upon query size, lengths of terms, wordiness, ambiguity, difficulty and objective; addressing each of these features requires specific approaches. Several experimental studies have also reported remarkable improvement in retrieval effectiveness: both with respect to precision and recall. These results are a proof of advancement of research of query expansion techniques. Based on recent trend in literature, hybrid techniques (combination of two or more techniques) give best results and seem to be more effective with respect to diversity of users, queries and document corpus.

With the ever growing wealth of information available on Internet, web searching has become an integral part of our lives. Every web user wants personalized information, and hence, information retrieval systems need to personalize search results based on the query and the user. We believe personalization of web search results will play an important in QE research in future.

References

1. Abdulla, A.A.A., Lin, H., Xu, B., Banbhrani, S.K.: Improving biomedical information retrieval by linear combinations of different query expansion techniques. BMC bioinformatics **17**(7), 238 (2016)
2. Adriani, M., Van Rijsbergen, C.: Term similarity-based query expansion for cross-language information retrieval. In: International Conference on Theory and Practice of Digital Libraries, pp. 311–322. Springer (1999)
3. Aggarwal, N., Buitelaar, P.: Query expansion using wikipedia and dbpedia. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
4. 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)
5. Agirre, E., de Lacalle, O.L., Soroa, A.: Random walks for knowledge-based word sense disambiguation. Computational Linguistics **40**(1), 57–84 (2014)
6. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Acm sigmod record, vol. 22, pp. 207–216. ACM (1993)
7. Al-Shboul, B., Myaeng, S.H.: Wikipedia-based query phrase expansion in patent class search. Information retrieval **17**(5-6), 430–451 (2014)
8. Allan, J.: Incremental relevance feedback for information filtering. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 270–278. ACM (1996)
9. 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)
10. ALMasri, M., Berrut, C., Chevallet, J.P.: A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information. In: European Conference on Information Retrieval, pp. 709–715. Springer (2016)
11. Amati, G., Joost, C., Rijsbergen, V.: Probabilistic models for information retrieval based on divergence from randomness (2003)
12. Amer, N.O., Mulhem, P., Géry, M.: Toward word embedding for personalized information retrieval. In: Neu-IR: The SIGIR 2016 Workshop on Neural Information Retrieval (2016)

13. Anand, R., Kotov, A.: An empirical comparison of statistical term association graphs with dbpedia and conceptnet for query expansion. In: Proceedings of the 7th Forum for Information Retrieval Evaluation, pp. 27–30. ACM (2015)
14. Arguello, J., Elsas, J.L., Callan, J., Carbonell, J.G.: Document representation and query expansion models for blog recommendation. ICWSM **2008**(0), 1 (2008)
15. Atefeh, F., Khreich, W.: A survey of techniques for event detection in twitter. Computational Intelligence **31**(1), 132–164 (2015)
16. Attar, R., Fraenkel, A.S.: Local feedback in full-text retrieval systems. Journal of the ACM (JACM) **24**(3), 397–417 (1977)
17. 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)
18. 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)
19. 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)
20. Bai, J., Nie, J.Y., Cao, G.: Context-dependent term relations for information retrieval. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 551–559. Association for Computational Linguistics (2006)
21. 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)
22. 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)
23. Ballesteros, L., Croft, B.: Dictionary methods for cross-lingual information retrieval. In: International Conference on Database and Expert Systems Applications, pp. 791–801. Springer (1996)
24. Ballesteros, L., Croft, W.B.: Phrasal translation and query expansion techniques for cross-language information retrieval. In: ACM SIGIR Forum, vol. 31, pp. 84–91. ACM (1997)
25. Ballesteros, L., Croft, W.B.: Resolving ambiguity for cross-language retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 64–71. ACM (1998)
26. Ballesteros, L.A.: Cross-language retrieval via transitive translation. In: Advances in information retrieval, pp. 203–234. Springer (2002)
27. Barrington, L., Chan, A., Turnbull, D., Lanckriet, G.: Audio information retrieval using semantic similarity. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, vol. 2, pp. II–725. IEEE (2007)
28. Belkin, N.J., Croft, W.B.: Information filtering and information retrieval: Two sides of the same coin? Communications of the ACM **35**(12), 29–38 (1992)
29. Bender, M., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J.X., Schenkel, R., Weikum, G.: Exploiting social relations for query expansion and result ranking. In: Data engineering workshop, 2008. ICDEW 2008. IEEE 24th International Conference on, pp. 501–506. IEEE (2008)
30. Bendersky, M., Metzler, D., Croft, W.B.: Parameterized concept weighting in verbose queries. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 605–614. ACM (2011)
31. Bernardini, A., Carpineto, C.: Fub at trec 2008 relevance feedback track: extending rocchio with distributional term analysis. Tech. rep., DTIC Document (2008)
32. 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)
33. Bhattacharya, P., Goyal, P., Sarkar, S.: Usingword embeddings for query translation for hindi to english cross language information retrieval. arXiv preprint arXiv:1608.01561 (2016)
34. Bhogal, J., Macfarlane, A., Smith, P.: A review of ontology based query expansion. Information processing & management **43**(4), 866–886 (2007)
35. Bian, J., Liu, Y., Agichtein, E., Zha, H.: Finding the right facts in the crowd: factoid question answering over social media. In: Proceedings of the 17th international conference on World Wide Web, pp. 467–476. ACM (2008)
36. Biancalana, C., Gasparetti, F., Micarelli, A., Sansonetti, G.: Social semantic query expansion. ACM Transactions on Intelligent Systems and Technology (TIST) **4**(4), 60 (2013)
37. Biancalana, C., Micarelli, A.: Social tagging in query expansion: A new way for personalized web search. In: Computational Science and Engineering, 2009. CSE'09. International Conference on, vol. 4, pp. 1060–1065. IEEE (2009)
38. 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)
39. Billerbeck, B., Zobel, J.: Questioning query expansion: An examination of behaviour and parameters. In: Proceedings of the 15th Australasian database conference-Volume 27, pp. 69–76. Australian Computer Society, Inc. (2004)
40. 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)
41. Boer, M., Schutte, K., Kraaij, W.: Knowledge based query expansion in complex multimedia event detection. Multimedia Tools and Applications pp. 1–19 (2015)

42. de Boer, M., Schutte, K., Kraaij, W.: Knowledge based query expansion in complex multimedia event detection. Multimedia Tools and Applications **75**(15), 9025–9043 (2016)
43. Borth, D., Ji, R., Chen, T., Breuel, T., Chang, S.F.: Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of the 21st ACM international conference on Multimedia, pp. 223–232. ACM (2013)
44. Bouadjenek, M.R., Hacid, H., Bouzeghoub, M.: Laicos: an open source platform for personalized social web search. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1446–1449. ACM (2013)
45. Bouadjenek, M.R., Hacid, H., Bouzeghoub, M.: Sopra: A new social personalized ranking function for improving web search. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp. 861–864. ACM (2013)
46. Bouadjenek, M.R., Hacid, H., Bouzeghoub, M., Daigremont, J.: Personalized social query expansion using social bookmarking systems. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 1113–1114. ACM (2011)
47. Bouadjenek, M.R., Hacid, H., Bouzeghoub, M., Vakali, A.: Persador: Personalized social document representation for improving web search. Information Sciences **369**, 614–633 (2016)
48. Bouchoucha, A., He, J., Nie, J.Y.: Diversified query expansion using conceptnet. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, pp. 1861–1864. ACM (2013)
49. Broder, A.: A taxonomy of web search. In: ACM Sigir forum, vol. 36, pp. 3–10. ACM (2002)
50. Buckley, C., Harman, D.: Reliable information access final workshop report. ARDA Northeast Regional Research Center Technical Report **3** (2004)
51. Buckley, C., Salton, G., Allan, J.: The effect of adding relevance information in a relevance feedback environment. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 292–300. Springer-Verlag New York, Inc. (1994)
52. Buckley, C., Salton, G., Allan, J., Singhal, A.: Automatic query expansion using smart: Trec 3. NIST special publication sp pp. 69–69 (1995)
53. Büttcher, S., Clarke, C.L., Cormack, G.V.: Information retrieval: Implementing and evaluating search engines. Mit Press (2016)
54. Cao, G., Gao, J., Nie, J.Y., Bai, J.: Extending query translation to cross-language query expansion with markov chain models. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 351–360. ACM (2007)
55. Cao, G., Nie, J.Y., Gao, J., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 243–250. ACM (2008)
56. 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)
57. 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)
58. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Computing Surveys (CSUR) **44**(1), 1 (2012)
59. Carpineto, C., Romano, G., Giannini, V.: Improving retrieval feedback with multiple term-ranking function combination. ACM Transactions on Information Systems (TOIS) **20**(3), 259–290 (2002)
60. Cavalin, P., Figueiredo, F., de Bayser, M., Moyano, L., Candello, H., Appel, A., Souza, R.: Building a question-answering corpus using social media and news articles. In: International Conference on Computational Processing of the Portuguese Language, pp. 353–358. Springer (2016)
61. Chang, Y., Ounis, I., Kim, M.: Query reformulation using automatically generated query concepts from a document space. Information processing & management **42**(2), 453–468 (2006)
62. Chirita, P.A., Firan, C.S., Nejdl, W.: Personalized query expansion for the web. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 7–14. ACM (2007)
63. Chu-Carroll, J., Prager, J., Czuba, K., Ferrucci, D., Duboue, P.: Semantic search via xml fragments: a high-precision approach to ir. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 445–452. ACM (2006)
64. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational linguistics **16**(1), 22–29 (1990)
65. Cilibrasi, R.L., Vitanyi, P.M.: The google similarity distance. IEEE Transactions on knowledge and data engineering **19**(3) (2007)
66. Collins-Thompson, K.: Reducing the risk of query expansion via robust constrained optimization. In: Proceedings of the 18th ACM conference on Information and knowledge management, pp. 837–846. ACM (2009)
67. 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)
68. Collins-Thompson, K., Callan, J.: Estimation and use of uncertainty in pseudo-relevance feedback. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 303–310. ACM (2007)
69. Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F., Voorhees, E.M.: Trec 2014 web track overview. Tech. rep., DTIC Document (2015)
70. Croft, B., Lafferty, J.: Language modeling for information retrieval, vol. 13. Springer Science & Business Media (2013)
71. Croft, W.B., Harper, D.J.: Using probabilistic models of document retrieval without relevance information. Journal of documentation **35**(4), 285–295 (1979)

72. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 299–306. ACM (2002)
73. 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)
74. 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)
75. 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)
76. Dalton, J., Dietz, L.: A neighborhood relevance model for entity linking. In: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, pp. 149–156. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE (2013)
77. 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)
78. Dang, E.K.F., Luk, R.W., Allan, J.: A context-dependent relevance model. Journal of the Association for Information Science and Technology **67**(3), 582–593 (2016)
79. 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)
80. Darwish, K., Magdy, W., et al.: Arabic information retrieval. Foundations and Trends® in Information Retrieval **7**(4), 239–342 (2014)
81. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR) **40**(2), 5 (2008)
82. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological) pp. 1–38 (1977)
83. Diaz, F.: Condensed list relevance models. In: Proceedings of the 2015 International Conference on The Theory of Information Retrieval, pp. 313–316. ACM (2015)
84. Diaz, F., Mitra, B., Craswell, N.: Query expansion with locally-trained word embeddings. arXiv preprint arXiv:1605.07891 (2016)
85. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
86. Doszkocs, T.E.: Aid, an associative interactive dictionary for online searching. Online Review **2**(2), 163–173 (1978)
87. Douze, M., Revaud, J., Schmid, C., Jégou, H.: Stable hyper-pooling and query expansion for event detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1825–1832 (2013)
88. Efthimiadis, E.N.: Query expansion. Annual review of information science and technology **31**, 121–187 (1996)
89. Egozi, O., Markovitch, S., Gabrilovich, E.: Concept-based information retrieval using explicit semantic analysis. ACM Transactions on Information Systems (TOIS) **29**(2), 8 (2011)
90. Eguchi, K.: Ntcir-5 query expansion experiments using term dependence models. In: NTCIR (2005)
91. Eichstaedt, M., Patel, A.P., Lu, Q., Manber, U., Rudkin, K.: System and method for personalized information filtering and alert generation (2002). US Patent 6,381,594
92. Fitzpatrick, L., Dent, M.: Automatic feedback using past queries: social searching? In: ACM SIGIR Forum, vol. 31, pp. 306–313. ACM (1997)
93. 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)
94. Franzoni, V.: Just an update on pming distance for web-based semantic similarity in artificial intelligence and data mining. arXiv preprint arXiv:1701.02163 (2017)
95. Franzoni, V., Milani, A.: Pming distance: A collaborative semantic proximity measure. In: Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 02, pp. 442–449. IEEE Computer Society (2012)
96. Fu, G., Jones, C.B., Abdelmoty, A.I.: Ontology-based spatial query expansion in information retrieval. In: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", pp. 1466–1482. Springer (2005)
97. Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T.: The vocabulary problem in human-system communication. Communications of the ACM **30**(11), 964–971 (1987)
98. Gaillard, B., Bouraoui, J.L., De Neef, E.G., Boualem, M.: Query expansion for cross language information retrieval improvement. In: RCIS, pp. 337–342 (2010)
99. Gan, L., Hong, H.: Improving query expansion for information retrieval using wikipedia. International Journal of Database Theory and Application **8**(3), 27–40 (2015)
100. Gao, Y., Xu, Y., Li, Y.: Pattern-based topics for document modelling in information filtering. IEEE Transactions on Knowledge and Data Engineering **27**(6), 1629–1642 (2015)
101. 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)
102. 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)
103. 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)
104. Gonzalo, J., Verdejo, F., Chugur, I., Cigarran, J.: Indexing with wordnet synsets can improve text retrieval. arXiv preprint cmp-lg/9808002 (1998)
105. Graupmann, J., Cai, J., Schenkel, R.: Automatic query refinement using mined semantic relations. In: Web Information Retrieval and Integration, 2005. WIRI'05. Proceedings. International Workshop on Challenges in, pp. 205–213. IEEE (2005)

106. Guisado-Gámez, J., Prat-Pérez, A., Larriba-Pey, J.L.: Query expansion via structural motifs in wikipedia graph. arXiv preprint arXiv:1602.07217 (2016)
107. Hahm, G.J., Yi, M.Y., Lee, J.H., Suh, H.W.: A personalized query expansion approach for engineering document retrieval. Advanced Engineering Informatics **28**(4), 344–359 (2014)
108. Hanani, U., Shapira, B., Shoval, P.: Information filtering: Overview of issues, research and systems. User modeling and user-adapted interaction **11**(3), 203–259 (2001)
109. Harman, D.: Relevance feedback and other query modification techniques. (1992)
110. Harman, D., Voorhees, E.: Overview of the seventh text retrieval conference (trec-7). In: Proceedings of the Seventh Text REtrieval Conference (TREC-7), NIST Special Publication, pp. 500–242 (1996)
111. He, B., Ounis, I.: Combining fields for query expansion and adaptive query expansion. Information processing & management **43**(5), 1294–1307 (2007)
112. He, B., Ounis, I.: Studying query expansion effectiveness. In: ECIR, vol. 9, pp. 611–619. Springer (2009)
113. Hersh, W., Price, S., Donohoe, L.: Assessing thesaurus-based query expansion using the umls metathesaurus. In: Proceedings of the AMIA Symposium, p. 344. American Medical Informatics Association (2000)
114. Hsu, M.H., Tsai, M.F., Chen, H.H.: Query expansion with conceptnet and wordnet: An intrinsic comparison. In: Asia Information Retrieval Symposium, pp. 1–13. Springer (2006)
115. Hsu, M.H., Tsai, M.F., Chen, H.H.: Combining wordnet and conceptnet for automatic query expansion: a learning approach. In: Asia information retrieval symposium, pp. 213–224. Springer (2008)
116. 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)
117. 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)
118. 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)
119. 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)
120. Huber, S., Seiger, R., Kühnert, A., Theodorou, V., Schlegel, T.: Goal-based semantic queries for dynamic processes in the internet of things. International Journal of Semantic Computing **10**(02), 269–293 (2016)
121. Huber, S., Seiger, R., Schlegel, T., et al.: Using semantic queries to enable dynamic service invocation for processes in the internet of things. In: 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), pp. 214–221. IEEE (2016)
122. Hull, D.A., et al.: Stemming algorithms: A case study for detailed evaluation. JASIS **47**(1), 70–84 (1996)
123. Imran, H., Sharan, A.: Selecting effective expansion terms for better information retrieval (2010)
124. Jaccard, P.: The distribution of the flora in the alpine zone. New phytologist **11**(2), 37–50 (1912)
125. Jamil, H.M., Jagadish, H.V.: A structured query model for the deep relational web. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1679–1682. ACM (2015)
126. Jammalamadaka, R.C., Salaka, V.K., Johnson, B.S., King, T.H.: Query expansion classifier for e-commerce (2015). US Patent 9,135,330
127. Jardine, N., van Rijsbergen, C.J.: The use of hierarchic clustering in information retrieval. Information storage and retrieval **7**(5), 217–240 (1971)
128. Jelinek, F.: Interpolated estimation of markov source parameters from sparse data. In: Proc. Workshop on Pattern Recognition in Practice, 1980 (1980)
129. Jelinek, F., Mercer, R.L.: Interpolated estimation of Markov source parameters from sparse data. In: Proceedings of the Workshop on Pattern Recognition in Practice (1980)
130. Jian, F., Huang, J.X., Zhao, J., He, T., Hu, P.: A simple enhancement for ad-hoc information retrieval via topic modelling. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 733–736. ACM (2016)
131. Jones, K.S.: Automatic keyword classification for information retrieval (1971)
132. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments: Part 2. Information processing & management **36**(6), 809–840 (2000)
133. Jones, S., Gatford, M., Robertson, S., Hancock-Beaulieu, M., Secker, J., Walker, S.: Interactive thesaurus navigation: intelligence rules ok? Journal of the American Society for Information Science **46**(1), 52 (1995)
134. Jourlin, P., Johnson, S.E., Jones, K.S., Woodland, P.C.: General query expansion techniques for spoken document retrieval. In: ESCA Tutorial and Research Workshop (ETRW) on Accessing Information in Spoken Audio (1999)
135. Junedi, M., Genevès, P., Layaïda, N.: Xml query-update independence analysis revisited. In: Proceedings of the 2012 ACM symposium on Document engineering, pp. 95–98. ACM (2012)
136. Kamps, J., Marx, M., Rijke, M.d., Sigurbjörnsson, B.: Articulating information needs in xml query languages. ACM Transactions on Information Systems (TOIS) **24**(4), 407–436 (2006)
137. Kang, I.H., Kim, G.: Query type classification for web document retrieval. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 64–71. ACM (2003)
138. 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)
139. Kato, M.P., Sakai, T., Tanaka, K.: Structured query suggestion for specialization and parallel movement: effect on search behaviors. In: Proceedings of the 21st international conference on World Wide Web, pp. 389–398. ACM (2012)
140. Khwileh, A., Jones, G.J.: Investigating segment-based query expansion for user-generated spoken content retrieval. In: Content-Based Multimedia Indexing (CBMI), 2016 14th International Workshop on, pp. 1–6. IEEE (2016)

141. Kim, Y., Seo, J., Croft, W.B.: Automatic boolean query suggestion for professional search. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 825–834. ACM (2011)
142. Kotov, A., Zhai, C.: Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries. In: Proceedings of the fifth ACM international conference on Web search and data mining, pp. 403–412. ACM (2012)
143. Kraaij, W., Nie, J.Y., Simard, M.: Embedding web-based statistical translation models in cross-language information retrieval. Computational Linguistics **29**(3), 381–419 (2003)
144. 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)
145. 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)
146. Krovetz, R., Croft, W.B.: Lexical ambiguity and information retrieval. ACM Transactions on Information Systems (TOIS) **10**(2), 115–141 (1992)
147. Kumar, N., Carterette, B.: Time based feedback and query expansion for twitter search. In: European Conference on Information Retrieval, pp. 734–737. Springer (2013)
148. Kuo, Y.H., Chen, K.T., Chiang, C.H., Hsu, W.H.: Query expansion for hash-based image object retrieval. In: Proceedings of the 17th ACM international conference on Multimedia, pp. 65–74. ACM (2009)
149. 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)
150. Lam-Adesina, A.M., Jones, G.J.: Applying summarization techniques for term selection in relevance feedback. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 1–9. ACM (2001)
151. Larkey, L.S., Ballesteros, L., Connell, M.E.: Light stemming for arabic information retrieval. In: Arabic computational morphology, pp. 221–243. Springer (2007)
152. Latiri, C., Haddad, H., Hamrouni, T.: Towards an effective automatic query expansion process using an association rule mining approach. Journal of Intelligent Information Systems **39**(1), 209–247 (2012)
153. Lau, R.Y., Bruza, P.D., Song, D.: Belief revision for adaptive information retrieval. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 130–137. ACM (2004)
154. Lau, T., Horvitz, E.: Patterns of search: analyzing and modeling web query refinement. In: UM99 User Modeling, pp. 119–128. Springer (1999)
155. Lavrenko, V., Allan, J.: Real-time query expansion in relevance models. Internal Report no 473, Center for Intelligent Information Retrieval-CIIR, University of Massachusetts (2006)
156. Lavrenko, V., Croft, W.B.: Relevance based language models. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 120–127. ACM (2001)
157. 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)
158. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al.: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web **6**(2), 167–195 (2015)
159. 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)
160. Levow, G.A., Oard, D.W., Resnik, P.: Dictionary-based techniques for cross-language information retrieval. Information processing & management **41**(3), 523–547 (2005)
161. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **2**(1), 1–19 (2006)
162. Li, X., Uricchio, T., Ballan, L., Bertini, M., Snoek, C.G., Bimbo, A.D.: Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. ACM Computing Surveys (CSUR) **49**(1), 14 (2016)
163. 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)
164. Lin, D., Pantel, P.: Discovery of inference rules for question-answering. Natural Language Engineering **7**(04), 343–360 (2001)
165. Lin, J., Murray, G.C.: Assessing the term independence assumption in blind relevance feedback. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 635–636. ACM (2005)
166. Liu, D., Yan, S., Ji, R.R., Hua, X.S., Zhang, H.J.: Image retrieval with query-adaptive hashing. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **9**(1), 2 (2013)
167. Liu, D.R., Chen, Y.H., Shen, M., Lu, P.J.: Complementary qa network analysis for qa retrieval in social question-answering websites. Journal of the Association for Information Science and Technology **66**(1), 99–116 (2015)
168. Liu, H., Singh, P.: Conceptneta practical commonsense reasoning tool-kit. BT technology journal **22**(4), 211–226 (2004)
169. 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)

170. Liu, X., Chen, F., Fang, H., Wang, M.: Exploiting entity relationship for query expansion in enterprise search. Information retrieval **17**(3), 265–294 (2014)
171. 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)
172. Lv, Y., Zhai, C.: Positional language models for information retrieval. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 299–306. ACM (2009)
173. Lv, Y., Zhai, C.: Positional relevance model for pseudo-relevance feedback. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 579–586. ACM (2010)
174. Lv, Y., Zhai, C., Chen, W.: A boosting approach to improving pseudo-relevance feedback. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 165–174. ACM (2011)
175. Magdy, W., Jones, G.J.: A study on query expansion methods for patent retrieval. In: Proceedings of the 4th workshop on Patent information retrieval, pp. 19–24. ACM (2011)
176. Mahdabi, P., Crestani, F.: The effect of citation analysis on query expansion for patent retrieval. Information Retrieval **17**(5-6), 412–429 (2014)
177. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
178. Maron, M.: Mechanized documentation: The logic behind a probabilistic. In: Statistical Association Methods for Mechanized Documentation: Symposium Proceedings, vol. 269, p. 9. US Government Printing Office (1965)
179. Maron, M.E., Kuhns, J.L.: On relevance, probabilistic indexing and information retrieval. Journal of the ACM (JACM) **7**(3), 216–244 (1960)
180. 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)
181. McNamee, P., Mayfield, J.: Comparing cross-language query expansion techniques by degrading translation resources. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 159–166. ACM (2002)
182. Metzler, D., Croft, W.B.: Latent concept expansion using markov random fields. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 311–318. ACM (2007)
183. Miao, J., Huang, J.X., Ye, Z.: Proximity-based rocchio's model for pseudo relevance. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 535–544. ACM (2012)
184. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
185. Mikroyannidis, A.: Toward a social semantic web. Computer **40**(11), 113–115 (2007)
186. 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)
187. 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)
188. 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)
189. Moldovan, D.I., Mihalcea, R.: Using wordnet and lexical operators to improve internet searches. IEEE Internet Computing **4**(1), 34 (2000)
190. Molino, P., Aiello, L.M., Lops, P.: Social question answering: Textual, user, and network features for best answer prediction. ACM Transactions on Information Systems (TOIS) **35**(1), 4 (2016)
191. Montague, M., Aslam, J.A.: Relevance score normalization for metasearch. In: Proceedings of the tenth international conference on Information and knowledge management, pp. 427–433. ACM (2001)
192. 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)
193. Mulhem, P., Amer, N.O., Géry, M.: Axiomatic Term-Based Personalized Query Expansion Using Bookmarking System, pp. 235–243. Springer International Publishing, Cham (2016)
194. 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)
195. Navigli, R.: Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) **41**(2), 10 (2009)
196. 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)
197. Nawab, R.M.A., Stevenson, M., Clough, P.: An ir-based approach utilising query expansion for plagiarism detection in medline (2016)
198. Nie, J.Y., Simard, M., Isabelle, P., Durand, R.: Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 74–81. ACM (1999)
199. Nie, L., Jiang, H., Ren, Z., Sun, Z., Li, X.: Query expansion based on crowd knowledge for code search. IEEE Transactions on Services Computing **9**(5), 771–783 (2016)
200. 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)
201. Paik, J.H., Pal, D., Parui, S.K.: Incremental blind feedback: An effective approach to automatic query expansion. ACM Transactions on Asian Language Information Processing (TALIP) **13**(3), 13 (2014)
202. Pakhomov, S.V., Finley, G., McEwan, R., Wang, Y., Melton, G.B.: Corpus domain effects on distributional semantic modeling of medical terms. Bioinformatics **32**(23), 3635–3644 (2016)

203. Pal, A.R., Saha, D.: Word sense disambiguation: A survey. arXiv preprint arXiv:1508.01346 (2015)
204. Pal, D., Mitra, M., Bhattacharya, S.: Exploring query categorisation for query expansion: A study. arXiv preprint arXiv:1509.05567 (2015)
205. Pal, D., Mitra, M., Datta, K.: Query expansion using term distribution and term association. arXiv preprint arXiv:1303.0667 (2013)
206. 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)
207. Pane, J.F., Myers, B.A.: Improving user performance on boolean queries. In: CHI'00 Extended Abstracts on Human Factors in Computing Systems, pp. 269–270. ACM (2000)
208. Panovich, K., Miller, R., Karger, D.: Tie strength in question & answer on social network sites. In: Proceedings of the ACM 2012 conference on computer supported cooperative work, pp. 1057–1066. ACM (2012)
209. Peat, H.J., Willett, P.: The limitations of term co-occurrence data for query expansion in document retrieval systems. Journal of the american society for information science **42**(5), 378 (1991)
210. Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K.: Dictionary-based cross-language information retrieval: Problems, methods, and research findings. Information retrieval **4**(3-4), 209–230 (2001)
211. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)
212. Porter, M.F.: Implementing a probabilistic information retrieval system. Information Technology: Research and Development **1**(2), 131–156 (1982)
213. Pound, J., Ilyas, I.F., Weddell, G.: Expressive and flexible access to web-extracted data: a keyword-based structured query language. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pp. 423–434. ACM (2010)
214. 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)
215. Radwan, K.: Vers l'acces multilingue en langage naturel aux bases de donnees textuelles. Ph.D. thesis (1994)
216. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007 (1995)
217. Riezler, S., Liu, Y., Vasserman, A.: Translating queries into snippets for improved query expansion. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, pp. 737–744. Association for Computational Linguistics (2008)
218. 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)
219. 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)
220. Rijsbergen, C.J.V.: Information Retrieval, 2nd edn. Butterworth-Heinemann, Newton, MA, USA (1979)
221. Rivas, A.R., Iglesias, E.L., Borrajo, L.: Study of query expansion techniques and their application in the biomedical information retrieval. The Scientific World Journal **2014** (2014)
222. Robertson, A.M., Willett, P.: A comparison of spelling-correction methods for the identification of word forms in historical text databases. Literary and linguistic computing **8**(3), 143–152 (1993)
223. Robertson, S.: Understanding inverse document frequency: on theoretical arguments for idf. Journal of documentation **60**(5), 503–520 (2004)
224. Robertson, S.E.: On term selection for query expansion. Journal of documentation **46**(4), 359–364 (1990)
225. Robertson, S.E., Jones, K.S.: Relevance weighting of search terms. Journal of the American Society for Information science **27**(3), 129–146 (1976)
226. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 232–241. Springer-Verlag New York, Inc. (1994)
227. Robertson, S.E., Walker, S.: Microsoft cambridge at trec-9: Filtering track. In: TREC (2000)
228. Robertson, S.E., Walker, S., Beaulieu, M., Willett, P.: Okapi at trec-7: automatic ad hoc, filtering, vlc and interactive track. Nist Special Publication SP (500), 253–264 (1999)
229. Rocchio, J.J.: Relevance feedback in information retrieval (1971)
230. Roy, D., Paul, D., Mitra, M., Garain, U.: Using word embeddings for automatic query expansion. arXiv preprint arXiv:1606.07608 (2016)
231. Ruthven, I., Lalmas, M.: A survey on the use of relevance feedback for information access systems. The Knowledge Engineering Review **18**(2), 95–145 (2003)
232. Salton, G.: Automatic text processing: The transformation, analysis, and retrieval of. Reading: Addison-Wesley (1989)
233. Salton, G.: Developments in automatic text retrieval. science **253**(5023), 974–980 (1991)
234. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management **24**(5), 513–523 (1988)
235. Salton, G., Buckley, C.: Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science **41**, 288–297 (1990)
236. Salton, G., Buckley, C.: Improving retrieval performance by relevance feedback. Readings in information retrieval **24**(5), 355–363 (1997)
237. Savoy, J.: Comparative study of monolingual and multilingual search models for use with asian languages. ACM transactions on Asian language information processing (TALIP) **4**(2), 163–189 (2005)
238. Shah, C., Croft, W.B.: Evaluating high accuracy retrieval techniques. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 2–9. ACM (2004)
239. Shekarpour, S., Höffner, K., Lehmann, J., Auer, S.: Keyword query expansion on linked data using linguistic and semantic features. In: Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on, pp. 191–197. IEEE (2013)

240. Sheridan, P., Ballerini, J.P.: Experiments in multilingual information retrieval using the spider system. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 58–65. ACM (1996)
241. Sihvonen, A., Vakkari, P.: Subject knowledge improves interactive query expansion assisted by a thesaurus. Journal of Documentation **60**(6), 673–690 (2004)
242. Singh, J., Prasad, M., Prasad, O.K., Joo, E.M., Saxena, A.K., Lin, C.T.: A novel fuzzy logic model for pseudo-relevance feedback-based query expansion. International Journal of Fuzzy Systems **18**(6), 980–989 (2016)
243. Singh, J., Sharan, A.: A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach. Neural Computing and Applications pp. 1–24 (2016)
244. Singhal, A., Pereira, F.: Document expansion for speech retrieval. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 34–41. ACM (1999)
245. 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)
246. Song, M., Song, I.Y., Hu, X., Allen, R.B.: Integration of association rules and ontologies for semantic query expansion. Data & Knowledge Engineering **63**(1), 63–75 (2007)
247. Song, R., Yu, L., Wen, J.R., Hon, H.W.: A proximity probabilistic model for information retrieval. Tech. rep., Tech. rep., Microsoft Research (2011)
248. Soricut, R., Brill, E.: Automatic question answering using the web: Beyond the factoid. Information Retrieval **9**(2), 191–206 (2006)
249. Spink, A., Wolfram, D., Jansen, M.B., Saracevic, T.: Searching the web: The public and their queries. Journal of the American society for information science and technology **52**(3), 226–234 (2001)
250. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 841–842. ACM (2010)
251. Stokoe, C., Oakes, M.P., Tait, J.: Word sense disambiguation in information retrieval revisited. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 159–166. ACM (2003)
252. 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)
253. Tejedor, J., Fapšo, M., Szöke, I., Černockỳ, J., Grézl, F., et al.: Comparison of methods for language-dependent and language-independent query-by-example spoken term detection. ACM Transactions on Information Systems (TOIS) **30**(3), 18 (2012)
254. Tellex, S., Kollar, T., Shaw, G., Roy, N., Roy, D.: Grounding spatial language for video search. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, p. 31. ACM (2010)
255. Thomas, S.S., Gupta, S., Venkatesh, K.: Perceptual synoptic view-based video retrieval using metadata. Signal, Image and Video Processing pp. 1–7 (2016)
256. Turtle, H.: Natural language vs. boolean query evaluation: A comparison of retrieval performance. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 212–220. Springer-Verlag New York, Inc. (1994)
257. 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)
258. Vaidya, J., Clifton, C.: Privacy preserving association rule mining in vertically partitioned data. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 639–644. ACM (2002)
259. Van Rijsbergen, C.J.: A non-classical logic for information retrieval. The computer journal **29**(6), 481–485 (1986)
260. Voorhees, E.M.: Query expansion using lexical-semantic relations. In: SIGIR94, pp. 61–69. Springer (1994)
261. Wang, F., Lin, L.: Domain lexicon-based query expansion for patent retrieval. In: Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on, pp. 1543–1547. IEEE (2016)
262. Wang, P., Xu, B., Xu, J., Tian, G., Liu, C.L., Hao, H.: Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. Neurocomputing **174**, 806–814 (2016)
263. 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)
264. 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)
265. Wei, X., Croft, W.B.: Modeling term associations for ad-hoc retrieval performance within language modeling framework. In: European Conference on Information Retrieval, pp. 52–63. Springer (2007)
266. Wen, J.R., Nie, J.Y., Zhang, H.J.: Query clustering using user logs. ACM Transactions on Information Systems **20**(1), 59–81 (2002)
267. 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)
268. Willett, P.: Recent trends in hierarchic document clustering: a critical review. Information Processing & Management **24**(5), 577–597 (1988)
269. Wong, W., Luk, R.W.P., Leong, H.V., Ho, K., Lee, D.L.: Re-examining the effects of adding relevance information in a relevance feedback environment. Information processing & management **44**(3), 1086–1116 (2008)

270. Wu, H., Fang, H.: An incremental approach to efficient pseudo-relevance feedback. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp. 553–562. ACM (2013)
271. 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)
272. Wu, J., Ilyas, I., Weddell, G.: A study of ontology-based query expansion. Technical report CS-2011–04 (2011)
273. Wu, Y., Liu, X., Xie, M., Ester, M., Yang, Q.: Cccf: Improving collaborative filtering via scalable user-item co-clustering. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, pp. 73–82. ACM (2016)
274. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp. 133–138. Association for Computational Linguistics (1994)
275. Xie, H., Zhang, Y., Tan, J., Guo, L., Li, J.: Contextual query expansion for image retrieval. IEEE Transactions on Multimedia **16**(4), 1104–1114 (2014)
276. 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)
277. 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)
278. Xu, J., Croft, W.B.: Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems (TOIS) **18**(1), 79–112 (2000)
279. 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)
280. Xue, G.R., Zeng, H.J., Chen, Z., Yu, Y., Ma, W.Y., Xi, W., Fan, W.: Optimizing web search using web click-through data. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management, pp. 118–126. ACM (2004)
281. Yao, Y., Yi, J., Liu, Y., Zhao, X., Sun, C.: Query processing based on associated semantic context inference. In: Information Science and Control Engineering (ICISCE), 2015 2nd International Conference on, pp. 395–399. IEEE (2015)
282. Yin, Z., Shokouhi, M., Craswell, N.: Query expansion using external evidence. In: European Conference on Information Retrieval, pp. 362–374. Springer (2009)
283. Yu, C.T., Buckley, C., Lam, K., Salton, G.: A generalized term dependence model in information retrieval. Tech. rep., Cornell University (1983)
284. Yu, K., Tresp, V., Yu, S.: A nonparametric hierarchical bayesian framework for information filtering. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 353–360. ACM (2004)
285. Zervakis, L., Tryfonopoulos, C., Skiadopoulos, S., Koubarakis, M.: Query reorganisation algorithms for efficient boolean information filtering. IEEE Transactions on Knowledge and Data Engineering (2016)
286. Zhai, C., Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of the tenth international conference on Information and knowledge management, pp. 403–410. ACM (2001)
287. Zhang, C.J., Zeng, A.: Behavior patterns of online users and the effect on information filtering. Physica A: Statistical Mechanics and its Applications **391**(4), 1822–1830 (2012)
288. 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)
289. Zhang, Y., Clark, S.: Syntactic processing using the generalized perceptron and beam search. Computational linguistics **37**(1), 105–151 (2011)
290. 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)
291. Zhong, Z., Ng, H.T.: Word sense disambiguation improves information retrieval. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pp. 273–282. Association for Computational Linguistics (2012)
292. Zhou, D., Lawless, S., Liu, J., Zhang, S., Xu, Y.: Query expansion for personalized cross-language information retrieval. In: Semantic and Social Media Adaptation and Personalization (SMAP), 2015 10th International Workshop on, pp. 1–5. IEEE (2015)
293. Zhou, D., Lawless, S., Wade, V.: Improving search via personalized query expansion using social media. Information retrieval **15**(3-4), 218–242 (2012)
294. Zhou, D., Lawless, S., Wu, X., Zhao, W., Liu, J., Lewandowski, D.: A study of user profile representation for personalized cross-language information retrieval. Aslib Journal of Information Management **68**(4) (2016)
295. Zhou, D., Wu, X., Zhao, W., Lawless, S., Liu, J.: Query expansion with enriched user profiles for personalized search utilizing folksonomy data. IEEE Transactions on Knowledge and Data Engineering (2017)
296. Zimmer, C., Tryfonopoulos, C., Weikum, G.: Exploiting correlated keywords to improve approximate information filtering. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 323–330. ACM (2008)
297. Zingla, M.A., Chiraz, L., Slimani, Y.: Short query expansion for microblog retrieval. Procedia Computer Science **96**, 225–234 (2016)