Information Processing and Management 56 (2019) 1698–1735

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Query expansion techniques for information retrieval: A survey

Hiteshwar Kumar Azad[*], Akshay Deepak

National Institute of Technology Patna, India

ARTICLE INFO

Keywords:
Query expansion
Query reformulation
Information retrieval
Internet search

ABSTRACT

With the ever increasing size of the web, relevant information extraction on the Internet with a query formed by a few keywords has become a big challenge. Query Expansion (QE) plays a crucial role in improving searches on the Internet. Here, the user's initial query is reformulated by adding additional meaningful terms with similar significance. QE – as part of information retrieval (IR) – has long attracted researchers' attention. It has become very influential in the field of personalized social document, question answering, cross-language IR, information filtering and multimedia IR. Research in QE has gained further prominence because of IR dedicated conferences such as TREC (Text Information Retrieval Conference) and CLEF (Conference and Labs of the Evaluation Forum). 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 – bringing out similarities and differences.

1. Introduction

There is a huge amount of data available on the Internet, and it is growing exponentially. This unconstrained information-growth has not been accompanied by a corresponding technical advancement in the approaches for extracting relevant information (Mikroyannidis, 2007). 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; as a result, the search results are not focused on the topic of interest. Second, the query can be too short to capture appropriately what the user is looking for. This can happen just as a matter of habit (e.g., the average size of a web search is 2.4 words Spink, Wolfram, Jansen, & Saracevic, 2001; Statista, 2017). Third, the user is often not sure about what he is looking for until he sees the results. Even if the user knows what he is searching for, he does not know how to formulate an appropriate query (navigational queries are exceptions to this Broder, 2002). QE plays an important part in fetching relevant results in the above cases.

Most web queries fall under the following three fundamental categories (Broder, 2002; Kang & Kim, 2003) :

- *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 are looking for specific website or URL (e.g., *ISRO*).
- *Transactional Queries:* Queries that demonstrate the user's intent to execute a specific activity (e.g., downloading papers or buying books).

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 queries and search index are not based on the same set of terms. This is also known as the vocabulary problem (Furnas, Landauer, Gomez, & Dumais, 1987); it results from a combination of synonymy and

[*] Corresponding author.
E-mail addresses: hiteshwar.cse15@nitp.ac.in (H.K. Azad), akshayd@nitp.ac.in (A. Deepak).

https://doi.org/10.1016/j.ipm.2019.05.009
Received 6 July 2018; Received in revised form 28 December 2018; Accepted 23 May 2019
Available online 28 May 2019

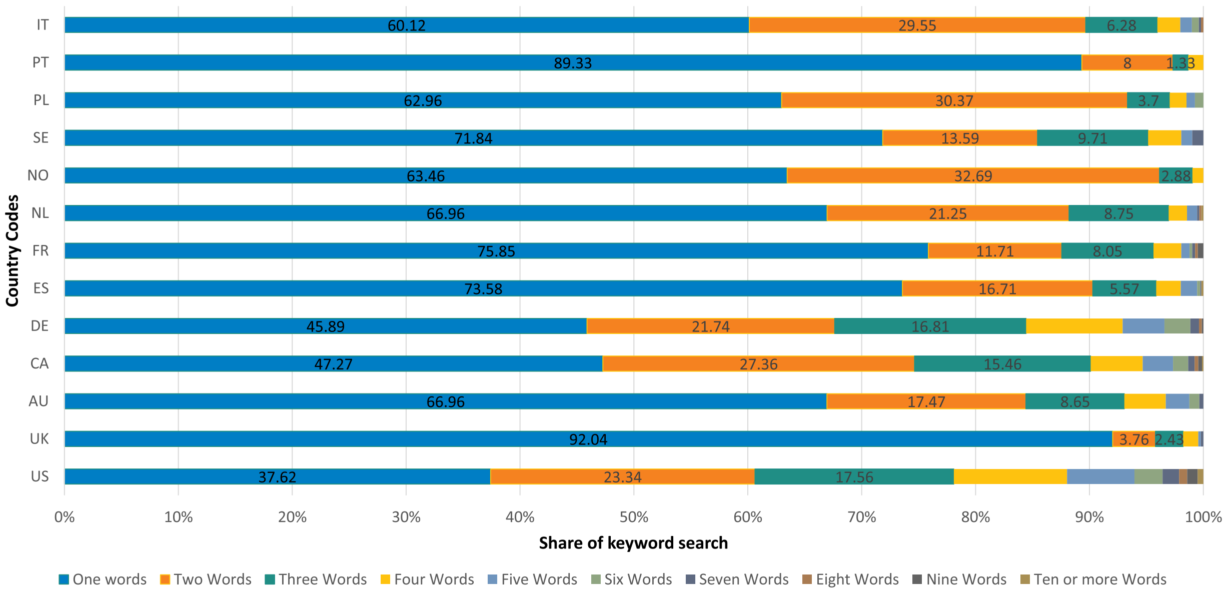

Fig. 1. Country-wise size of query searched on Internet constructed using data from Keyword (2018).

polysemy. Synonymy refers to multiple words with common meaning, e.g., “buy” and “purchase”. Polysemy refers to words with multiple meanings, e.g., “mouse” (a computer device or an animal). Synonymous and polysemous words are hindrances in retrieving relevant information; they reduce recall and precision rates.

To address the vocabulary problem, various techniques have been proposed, such as, relevance feedback, interactive query filtration, corpus dependent 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 applied by Maron and Kuhns (1960) as a technique for literature indexing and searching in a mechanized library system. It was Rocchio (1971) who brought QE to spotlight through “relevance feedback” and its characterization in a vector space model. The idea behind relevance feedback is to incorporate the 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 the relevance of the results. Rocchio’s work was further extended and applied in techniques such as collection-based term co-occurrence (Jones, 1971; van Rijsbergen, 1977), cluster-based information retrieval (Jardine & van Rijsbergen, 1971; Minker, Wilson, & Zimmerman, 1972), comparative analysis of term distribution (Porter, 1982; Van Rijsbergen, 1986; Yu, Buckley, Lam, & Salton, 1983) and automatic text processing (Salton, 1989; 1991; Salton & Buckley, 1988).

The above was before the search engine era, where search-retrieval was done on a small amount of data with short queries and satisfactory results were also obtained. In the 1990s, search engines were introduced, and suddenly, huge amounts of data started being published on the web, which has continued to grow at an exponential rate since then. However, users continued to fire short queries for web searches. While the recall rate suddenly increased, there was a loss in precision (Harman, 1992; Salton & Buckley, 1990). This called for modernization of QE techniques to deal with Internet-data.

As per recent reports (Keyword, 2018; Statista, 2017), the most frequent queries consist of one, two or three words only (see Fig. 1) – the same as seventeen years ago as reported by Lau and Horvitz (1999). While the query terms have remained few, the number of web pages have increased exponentially. This has increased the ambiguity – caused due to the multiple meanings/senses of the query terms (also called vocabulary mismatch problem) – in finding relevant pages. Hence, the importance of QE techniques has also increased in resolving the vocabulary mismatch problem.

Recently, QE has come to spotlight because a lot of researchers are using QE techniques for working on personalized social bookmarking services (Biancalana, Gasparetti, Micarelli, & Sansonetti, 2013; Bouadjenek, Hacid, Bouzeghoub, & Vakali, 2016; Ghorab, Zhou, O’Connor, & Wade, 2013), Question Answering over Linked Data (QALD)[1] (Unger, Ngomo, & Cabrio, 2016) and in Text Retrieval Conference (TREC).[2] QE techniques are also used heavily in web, desktop and email searches (Pal, Mitra, & Bhattacharya, 2015). Many platforms provide QE facility to end users, which can be turned on or off, e.g., WordNet,[3] ConceptNet,[4]

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

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

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

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

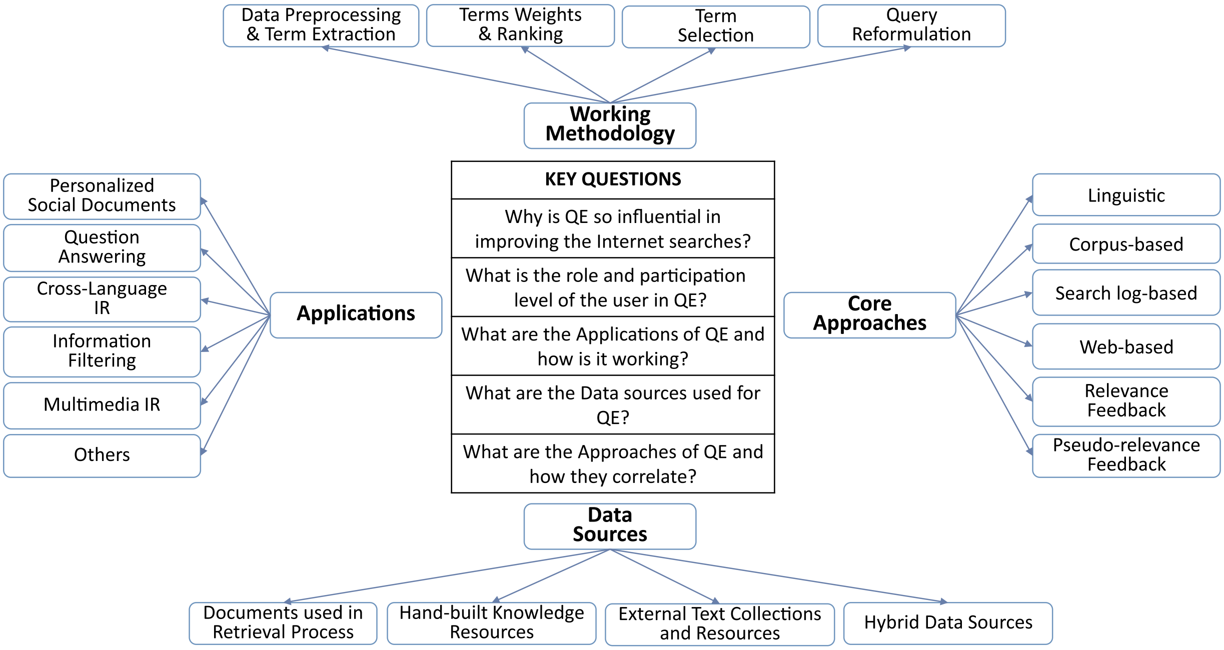
Data Preprocessing & Term Extraction
Terms Weights & Ranking
Term Selection
Query Reformulation
Working Methodology
KEY QUESTIONS
Why is QE so influential in improving the Internet searches?
What is the role and participation level of the user in QE?
What are the Applications of QE and how is it working?
What are the Data sources used for QE?
What are the Approaches of QE and how they correlate?
Personalized Social Documents
Question Answering
Cross-Language IR
Information Filtering
Multimedia IR
Others
Applications
Core Approaches
Linguistic
Corpus-based
Search log-based
Web-based
Relevance Feedback
Pseudo-relevance Feedback
Data Sources
Documents used in Retrieval Process
Hand-built Knowledge Resources
External Text Collections and Resources
Hybrid Data Sources

Fig. 2. Survey overview.

Lucene,[5] Google Enterprise[6] and MySQL.[7]

However, there are also drawbacks of QE techniques, e.g., there is a computational cost associated with the application of QE techniques. In the case of Internet searches, where quick response time is a must, the computational cost associated with the application of QE techniques prohibits their use in part or entirety (Imran & Sharan, 2010). Another drawback is that sometimes it can fail to establish a relationship between a word in the corpus with those being used in different communities, e.g., “senior citizen” and “elderly” (Gauch, Wang, & Rachakonda, 1999). Another issue is that QE may hurt the retrieval effectiveness for some queries (Collins-Thompson, 2009; Lv, Zhai, & Chen, 2011).

Few surveys have been done in the past on QE techniques. In 2007, Bhogal, Macfarlane, and Smith (2007) reviewed ontology-based QE techniques, which are domain specific. Such techniques have also been described in book by Manning, Raghavan, and Schütze (2008). Carpineto and Romano (2012) reviewed the major QE techniques, data sources, and features in an IR system. However, their survey covers only automatic query expansion (AQE) techniques and does not include recent research on personalized social documents, term weighting and ranking methods, and categorization of several data sources. After this, we could not find any significant review covering recent progress in QE techniques. 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 as summarized in Fig. 2.

The rest of the article is organized as follows. Section 2 defines QE and describes the working methodology of QE and outlines the main steps. Section 3 discuss the importance and application of QE. It also briefly discusses several applications of QE including those in recent literature. Section 4 classifies the existing approaches on the basis of properties of various data sources, and comparative analysis of these QE approaches. Finally, Section 5 discuss recent trends in literature and concludes the paper.

2. Query expansion

Query expansion reformulates the 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:

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

In the above definition, the key aspect of QE is the set T': set of new meaningful terms added to the user’s original query in order to retrieve more relevant documents and reduce ambiguity. Krovetz and Croft (1992) reported that this set T' computed on the basis of term similarity, and without changing the concept, increases recall rate in query results. Hence, computation of set T' and choice of

[5] http://lucene.apache.org/ .
[6] https://enterprise.google.com/search/ .
[7] https://www.mysql.com/ .

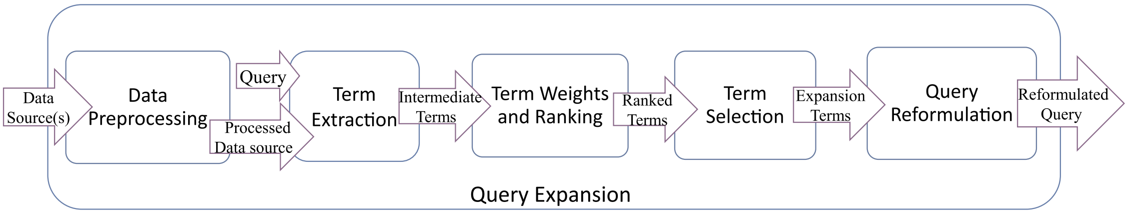

Fig. 3. Query expansion process.

data sources D are key aspects of research in QE.

In regard to automation and the end-user involvement (Efthimiadis, 1996), QE techniques can be classified as follows:

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

Query expansion working methodology:

The process of expanding query consists of four steps: (i) preprocessing of data sources and term extraction, (ii) term weights and ranking, (iii) term selection, and (iv) query reformulation (see Fig. 3). These steps are discussed next.

2.1. Preprocessing of data sources and term extraction

Preprocessing of a data source depends upon the data source and the approach being used for QE; it does not depend on the user's query. The primary goal of this step is to extract a set of terms from the data source that meaningfully augment the user's original query. It consists of the following four sub-steps:

1. Text extraction from the data source (extraction of whole texts from the specific data source used for QE)
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 of reducing derived or inflected words to their base word)

After preprocessing the raw data sources, the combined processed data source and the user query are used for term extraction.

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

2.1.1. Documents used in retrieval process

At the beginning of the seventies, the addition of similar terms into the initial query started playing a crucial role in QE (e.g., Jardine & van Rijsbergen, 1971; Minker et al., 1972; Willett, 1988). Researchers assumed that a set of similar words that frequently appear in documents, belong to the same subject, thus, similar documents formed a cluster (Peat & Willett, 1991). Two types of clustering have been discussed in document retrieval systems: clustering of terms and clustering of documents (Willett, 1988). A well-known example of term based clustering is Qiu and Frei (1993)'s corpus-based expansion technique that uses a 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. Another approach proposed by Crouch and Yang (1992) built a statistical corpus thesaurus by clustering the entire document collection using the link clustering algorithm. Some other works that use collection-based data sources for QE are Attar and Fraenkel (1977); Carpineto, De Mori, Romano, and Bigi (2001); Gauch et al. (1999); Jones (1971); Xu and Croft (1996) and Bai, Song, Bruza, Nie, and Cao (2005). Carpineto et al. (2001) used corpus as data sources from top retrieved documents on the basis of term co-occurrence in the entire document collection. Similarly, Bai et al. (2005) use collection-based data sources as the top-ranked documents and chooses the expansion terms on the basis of term co-occurrence and information flow over the entire corpus. Recently, Zhang, Wang, Si, and Gao (2016) used four corpora as data sources (one industry and three academic corpora) and presented a Two-stage Feature Selection framework (TFS) for query expansion known as the Supervised Query Expansion (SQE). The first stage is an 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 the 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 QE is that they fail to establish a relationship between a word in the corpus and those which are used in different communities, e.g., "senior citizen" and "elderly" (Gauch et al., 1999).

2.1.2. Hand-built knowledge resources

The primary goal of hand-built knowledge resources is to extract knowledge from textual hand-built data sources such as dictionaries, thesaurus, ontologies, Wikipedia and LOD cloud. Thesaurus-based QE can be either automatic or hand-built. One of the famous hand-built thesaurus is WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). Voorhees (1994) utilized WordNet to expand the original query with semantically similar terms called synsets. It was observed that the retrieval effectiveness improved significantly for unstructured queries, while only marginal improvement was observed for structured queries. Some other articles have also used WordNet to expand the original query, for example, Smeaton, Kelledy, and O'Donnell (1995) use synsets of the initial query and assign half weight, Liu, Liu, Yu, and Meng (2004) use word sense, Gong, Cheang et al. (2006) use semantic similarity, Zhang, Deng, and Li (2009) use concepts and Pal, Mitra, and Datta (2014) use semantic relations from WordNet. Pal et al. (2014) propose a new and effective way of using WordNet for QE, where Candidate Expansion Terms (CET) are selected from a set of pseudo-relevant documents and the usefulness of these terms is determined by considering multiple sources of information. The semantic relation between the expanded terms and the query terms is determined using WordNet. Lemos, de Paula, Zanichelli, and Lopes (2014) present an automatic query expansion (AQE) approach that uses word relations to increase the chances of finding relevant code. As data sources, it uses a thesaurus containing only software-related word relations and WordNet for expanding the user's query. Similarly, Hsu, Tsai, and Chen (2006) use ConceptNet (Liu & Singh, 2004) (having higher concept diversity) and WordNet (having higher discrimination ability) as the data sources for expanding the user's query. ConceptNet is a relational semantic network that helps to understand the common sense knowledge of texts written by users. Recently, a number of researchers used ConceptNet as the data source for QE (e.g., Anand & Kotov, 2015; Bouchoucha, He, & Nie, 2013; Hsu, Tsai, & Chen, 2008; Kotov & Zhai, 2012. Bouchoucha et al. (2013) use ConceptNet for QE and propose 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 are different from the previously selected expansion terms. Then, the top N expansion terms having the highest MMRE scores are selected. Recently, Wikipedia and DBpedia are being used widely as data sources for QE (e.g., Aggarwal & Buitelaar, 2012; ALMasri, Berrut, & Chevallet, 2013; Anand & Kotov, 2015; Arguello, Elsas, Callan, & Carbonell, 2008; Guisado-Gámez, Prat-Pérez, & Larriba-Pey, 2016; Li, Luk, Ho, & Chung, 2007; Xu, Jones, & Wang, 2009). Li et al. (2007) performed an investigation using Wikipedia and retrieved all articles corresponding to the original query as a source of expansion terms for pseudo-relevance feedback. It observed that for a particular query where the general pseudo-relevance feedback fails to improve the query, Wikipedia-based pseudo-relevance feedback improves it significantly. Xu et al. (2009) 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 sense that cover a narrow topic) and (3) broader queries (queries having neither ambiguous nor specific meaning). They consolidated the expansion terms into the original query and evaluated these techniques using language modeling IR. ALMasri et al. (2013) use Wikipedia for semantic enrichment of short queries based on in-link and out-link articles.

Augenstein, Gentile, Norton, Zhang, and Ciravegna (2013) use LOD cloud for keyword mapping and exploits the graph structure within the Linked Data to determine relations between resources that are useful to discover or to express semantic similarity directly. Utilization of data sources as knowledge bases in IR 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 (consisting of entities, their attributes, and their relationships to other entities) are quite popular as data sources for QE. Recently, Xiong and Callan (2015) use the knowledge base "freebase" (a large public knowledge base that contains semi-structured information about real-world entities and their facts) for improving QE. For the selection of the expansion terms, Xiong and Callan (2015) developed two methods: (1) utilization of tf-idf based Pseudo-Relevance Feedback on the linked objects' descriptions, and (2) utilization of Freebase's entity categories, which grant an ontology tree that illustrates entities at several levels of abstraction.

However, Hersh, Price, and Donohoe (2000) used a thesaurus relationship for QE in the UMLS Metathesaurus; they reported that nearly all types of QE reduce the recall and precision based on retrieval effectiveness. In their result, not surprisingly, 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 a proper noun and they have to be kept up to date. Experiments with QE using hand-built knowledge resources do not always show improvements in retrieval effectiveness. It does not improve well-formulated user queries, but significantly improves the retrieval effectiveness of poorly constructed queries.

2.1.3. External text collections and resources

External text collections (such as the WWW, Anchor text, Query logs, External corpus) used in the retrieval process are the most common and useful data sources for QE. In such cases, QE approaches show overall better performance in comparison to all the discussed data sources. Some data sources under this category need preprocessing procedures for text collection. For example, Kraft and Zien (2004), and, Dang and Croft (2010) use the anchor texts as the data source; they parse 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 suggest that anchor texts can be used to improve the traditional QE based on query logs. Click through records (URLs, queries) extracted from a search engine (Query logs) is another data source for QE, where users' queries are expanded based on correlation between the query terms and the document terms determined using user logs (e.g., Cui, Wen, Nie, & Ma, 2003; Wen, Nie, & Zhang, 2002). Some researchers refer to query logs as user logs since they are derived from historical records of user queries registered in the query logs of search engines (e.g., Baeza-Yates, Hurtado, & Mendoza, 2004; Billerbeck, Scholer, Williams, & Zobel, 2003; Cui, Wen, Nie, & Ma, 2002; Yin, Shokouhi, & Craswell, 2009). Yin et al. (2009) express the search engine query log as a bipartite graph, where query nodes are connected to the URL nodes by click edges; they reported an improvement of retrieval effectiveness by more than 10% in average precision. Wang and Zhai (2008) use web corpus and training data as data sources, and then extract query terms using search logs. Most of the search engines and related surveyed papers using QE are based on query logs. However, for customized search systems for Internet search, enterprise search, personalized search (such as the desktop or email search), or for infrequent queries, query logs are either not available or the user's past queries are not sufficient to describe the information needed. To overcome this limitation, Bhatia, Majumdar, and Mitra (2011) proposed a document-centric probabilistic model to generate query suggestions from the corpus that does not depend on query logs and utilizes only the co-occurrence of terms in the corpus. Besides extracting from the user logs, some of the researchers use the sequence of characters comprising the user's query and the corresponding documents from user clicks on URL. This may be useful to remove unwanted content and to find semantically similar terms (Beeferman & Berger, 2000).

Today, word embedding techniques are widely used for QE. Recently, Roy, Paul, Mitra, and Garain (2016) proposed a word embedding framework based on distributed neural language model word2vec. Based on the framework, it extracted similar terms to a query using the 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 QE methods perform more or less the same with and without any feedback information. Some other works using word embedding techniques are Diaz, Mitra, and Craswell (2016) and Kuzi, Shtok, and Kurland (2016). Diaz et al. (2016) presented a QE technique based on locally-trained word embedding (such as word2vec and GloVe) for ad hoc IR. They also used local embeddings that capture the nuances of topic-specific languages and are better than global embeddings. They also suggested that embeddings be learned on topically-constrained corpora, instead of large topically-unconstrained corpora. In a query-specific manner, their experimental results suggest towards adopting local embeddings instead of global embedding because of formers potentially superior representation. Similarly, Kuzi et al. (2016) proposed a QE technique based on word embeddings that uses Word2Vec's Continuous Bag-of-Words (CBOW) approach (Mikolov, Chen, Corrado, & Dean, 2013); 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. Singh and Sharan (2016) and Singh et al. (2016) used a fuzzy logic-based QE technique, and, the top-retrieved documents (obtained using pseudo-relevance feedback) as data sources. Here, each expansion term (obtained from the 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.

2.1.4. Hybrid data sources

Hybrid data sources are a combination of two or more data sources (such as the combination of (1) Document used in retrieval process, (2) hand-built knowledge resources, and (3) External text collection and resources). A good number of published works have used hybrid data sources for QE. For example, Collins-Thompson and Callan (2005) use a combination of query-specific term dependencies from multiple sources such as WordNet, an external corpus, and the top retrieved documents as data sources. He and Ounis (2007) use a combination of anchor text, top retrieve documents and corpus as data sources for QE. The main focus is to improve the quality of query term reweighting – rather than choosing the best terms – by taking a linear combination of the term frequencies of anchor text, title and body in the retrieved documents. Recently, Pal, Mitra, and Datta (2013) used data sources based on term distributions (using Kullback–Leibler Divergence (KLD) and Bose–Einstein statistics (Bo1)) and term association (using Local Context Analysis (LCA) and Relevance-based Language Model (RM3)) methods for QE. The experimental result demonstrated that the combined method gives better result in comparison to each individual method. Other research works based on hybrid resources are Lee, Croft, and Allan (2008); Wu et al. (2014) and Dalton, Dietz, and Allan (2014). Wu et al. (2014) use a hybrid data source, which is a combination of three different sources, namely community question answering (CQA) archive, query logs, and web search results. Different types of sources provide different types of signals and reveal the user's intentions from different perspectives. From web search logs they gain an understanding of the wider preference of common web users, from question descriptions they obtain some specific and question-oriented intent, and from the top web search results they further extract some of the popular topics related to the short queries. Dalton et al. (2014) 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 Anand and Kotov (2015), the document collection and external resources (encyclopedias such as DBpedia and knowledge bases such as ConceptNet) are the data sources for QE. 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.

Comparative analysis: In all the previously discussed data sources, hybrid data sources have been widely used for QE, hence, they can be considered as state-of-art. The main reason behind their widespread acceptance is that they include various features of the user's queries, which cannot be considered by any of the individual data sources. In research involving hybrid data sources, Wikipedia is a popular data source because it is freely available and is the largest encyclopedia on the web, where articles are regularly updated and new articles are added. However, Wikipedia shows good retrieval effectiveness for short queries only. Data sources belonging to the documents used in the retrieval process have a drawback that they fail to establish a relationship between a word used in a corpus to words used in the other corpora (e.g., "senior citizen" and "elderly"). In hand-built knowledge resources, it has been observed that the retrieval effectiveness improved significantly for unstructured queries, while only marginal improvement has been found for structured queries. Mainly, there are three limitations in hand-built knowledge resources: they are usually domain specific, typically do not contain a proper noun and they should be kept up to date. External text collection and resources show

Table 1
Summary of research in classification of data sources used in QE.

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 van Rijsbergen (1971), Minker et al. (1972), Willett (1988)
	Corpus or collection based data sources	Terms collection from specific domain knowledge	Jones (1971), Attar and Fraenkel (1977), Peat and Willett (1991), Crouch and Yang (1992), Qiu and Frei (1993), Xu and Croft (1996), Gauch et al. (1999), Carpineto et al. (2001), Bai et al. (2005)
Hand built knowledge resources	WordNet & Thesaurus	Word sense and synset	Miller et al. (1990), Voorhees (1994), Smeaton et al. (1995), Liu et al. (2004), Gong et al. (2006), Zhang et al. (2009), Pal et al. (2014)
	ConceptNet & Knowledge bases	Common sense knowledge and Freebase	Liu and Singh (2004), Hsu et al. (2006), Hsu et al. (2008), Kotov and Zhai (2012), Bouadjenek et al. (2013b), Anand and Kotov (2015)
	Wikipedia or DBpedia	Articles, titles & hyper links	Li et al. (2007), Arguello et al. (2008), Xu et al. (2009), Aggarwal and Buitelaar (2012), ALMasri et al. (2013), Al-Shboul and Myaeng (2014), Anand and Kotov (2015), Guisado-Gámez et al. (2016)
External text collections and resources	Anchor texts	Adjacent terms in anchor text or text extraction from anchor tags	Kraft and Zien (2004), Dang and Croft (2010)
	Query logs or user logs	Historical records of user queries registered in the query logs of search engine	Wen et al. (2002), Cui et al. (2003), Billerbeck et al. (2003), Baeza-Yates et al. (2004), Yin et al. (2009), Hua et al. (2013)
	External corpus	Nearby terms in word embedding framework	Roy et al. (2016), Diaz et al. (2016), Kuzi et al. (2016), Beeferman and Berger (2000)
Hybrid data sources	Top-ranked documents & multiple sources	All terms in top retrieved documents	Collins-Thompson and Callan (2005), He and Ounis (2007), Lee et al. (2008), Pal et al. (2013), Wu et al. (2014), Dalton et al. (2014), Singh and Sharan (2016)

overall better performance in comparison to the first two sources discussed earlier. However, some data sources under this category need preprocessing procedure for text collection (e.g., anchor text and query logs). In the case of query logs, it is possible that the query logs are either not available or the user's past queries are not sufficient to describe the information need.

Table 1 summarizes the classification of Data Sources used in QE in literature based on the above discussion.

2.2. Weighting and ranking of query expansion terms

In this step of QE, weights and ranks are assigned to query expansion terms obtained after data preprocessing (see Fig. 3). The input to this step is the user's query and texts extracted from the data sources in the first step. Assigned weights denote relevancy of the 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. Carpineto and Romano (2012) classify the techniques into four categories on the basis of a relationship between the query terms and the expansion features:

- *One-to-one association:* Correlates each expanded term to at least one query term.
- *One-to-many association:* Correlates each expanded term to many 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 chooses expansion terms having the highest probability.

The first two approaches can also be considered as local techniques. These are based on association hypothesis projected by Rijsbergen (1979): "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 is primarily motivated by Maron (1965). Rijsbergen (1979) outlines this concept as "to enlarge the initial request by using additional index terms that have a similar or related meaning to those of the given request". The above approaches have been discussed next.

2.2.1. One-to-one association

Weighting and ranking the expansion terms based on one-to-one association between the query terms and expansion terms is the most common approach for doing so. Here, each expansion term is correlated to (at least) one query term (hence the name "one-to-one"). Weights are assigned to each query term using one of the several techniques described next.

A popular approach to establish a one-to-one association is to use linguistic associations, namely, stemming algorithm. It is used 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 (Porter, 1980), 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 (Voorhees, 1994). Using WordNet, each query term is mapped to its synonyms and a similar set of words – obtained from WordNet – in the expanded query. For example, if we consider word “java” as a noun in WordNet, there are three synsets with 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 followed using ConceptNet (Hsu et al., 2006) to find the related concepts of a user’s queries for query expansion. For example, the word “file” has a related concept namely “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 the 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 set of terms. There are several approaches to determine term similarity.

Path length-based measures determine the term similarity between the synsets (senses) – obtained from WordNet – based on the path length of the linked synsets. Generally, path length-based measures include two similarity measurement techniques: shortest path similarity (Resnik, 1995) and Wu & Palmer (WP) similarity score (Wu & Palmer, 1994). 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. Then, the Shortest path similarity score (Resnik, 1995) is defined 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 value of the similarity score can be 1.

The Wu & Palmer (WP) similarity score (Wu & Palmer, 1994) 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

$d(s_1)$ ($d(s_2)$) is the depth of sense s_1 (s_2) from the root node R in WordNet (see Fig. 4).

Similarity score of WP similarity varies from 0 to 1 (precisely, $0 < Sim_{WP} \leq 1$).

Other approaches like the Jaccard coefficient and Dice coefficient are also used widely for similarity measurement. The Jaccard coefficient (Jaccard, 1912) 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 (Dice, 1945) is described as

$$Sim_{Dice}(s_1, s_2) = \frac{2.\ df_{s_1 \wedge s_2}}{df_{s_1} + dfs_2} \tag{5}$$

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

Term-document matrix is a two dimensional matrix M, whose rows represent the terms and columns represent the documents. Cell $M_{t,d}$ contains value $w_{t,\ d}$, where $w_{t,\ d}$ denotes the weight of term t in document d. Term-document matrix is used to compute the similarity score through correlation matrix $C = MM^T$, where each cell c_{s_1,s_2} denotes correlation (similarity) score between terms s_1 and s_2, and is defined as:

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

where $w_{s_1,j}$ ($w_{s_2,j}$) is the weight of term s_1 (s_2) in jth document.

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

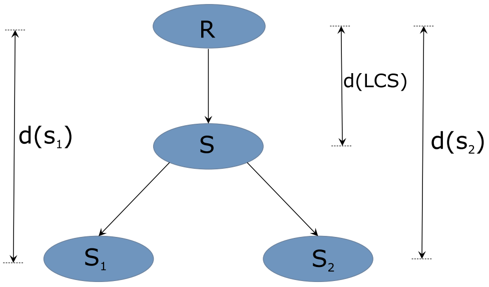

Fig. 4. Example of taxonomy hierarchy in WordNet.

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

where:

c_{s_1,s_2} denotes correlation (similarity) score between terms s_1 and s_2, and

$w_{s_1,j}$ ($w_{s_2,j}$) is the weight of term s_1 (s_2) in jth document.

Normalization is done to account for the relative frequency of terms.

It can be seen that using Eq. (6) we can create a set of conceptually different term-to-term correlation methods 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 a document.

A more exhaustive measurement technique for term co-occurrence that includes term dependency is mutual information (Church & Hanks, 1990):

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

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 probabilities of occurrence of terms s_1 and s_2.

These probabilities can be evaluated by relative frequency count. Eq. (8) is based on mutual information, which is symmetric in nature, i.e., $I(s_1, s_2) = I(s_2, s_1)$. However, in the context of words, order is important, i.e., "program executing" is different from "executing program". Hence, it is preferable to consider an asymmetric version of Eq. (8), where $P(s_1, s_2)$ refers to the probability that s_1 exactly follows s_2. The mutual information (Eq. (8)) will be: zero if there is no co-occurrence, one if terms s_1 and s_2 are distinct, and, $log_2(\frac{1}{P(s_1)} + 1)$ if s_1 is completely 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 (Bai et al., 2005) is identical to the association rule used in data mining problem (Agrawal, Imieliński, & Swami, 1993; Vaidya & Clifton, 2002). Association rules have been used widely for identifying the expansion feature correlation with the user query terms (Latiri, Haddad, & Hamrouni, 2012; Song, Song, Hu, & Allen, 2007).

Another corpus-based term similarity measure based on information content-based measurement is Resnik similarity (Resnik, 1995). 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 or equal to zero. The Resnik similarity is formulated as:

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

where $-\ log\, p(LCS(s_1, s_2))$ is the information content of the closest common ancestor node of two synsets s_1 and s_2.

The information content of a synset is defined as the logarithm of the probability of finding the synset in a given corpus. The negative sign makes the similarity score positive because probabilities are always between [0,1].

Recently, Wikipedia has become popular for expansion of short queries. In Wikipedia, it is possible 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, the term "apple" has two articles in Wikipedia, one indicating it as a fruit and the other as a company. ALMasri et al. (2013) use Wikipedia for semantic enhancement of short queries and measure 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) as in-links and

$O(s_1)$ ($O(s_2)$) is the set of articles that s_1 (s_2) points to as out-links.

Table 2 summarizes the mathematical form of term similarity score in One-to-One association based on the above discussion.

2.2.2. One-to-many association

In one-to-one association, a candidate term is added to the expanded query if it is correlated to even one term from the original query. The main issue with one-to-one association is that it may not properly demonstrate the connectivity between the expansion term and the query as a whole. For example, consider queries "data technology" and "music technology". Here, the word "technology" is frequently associated with the word "information". Hence, for query "data technology", a one-to-one association based

Table 2
Summery of one-to-one association for term ranking based on the term similarity score.

Reference	Approaches	Mathematical form
Jaccard (1912)	Jaccard coefficient	$\frac{df_{s1 \wedge s2}}{df_{s1 \vee s2}}$
Dice (1945)	Dice coefficient	$\frac{2\,.\, df_{s1 \wedge s2}}{df_{s_1} + dfs_2}$
Attar and Fraenkel (1977)	Cosine similarity	$\frac{\sum_{d_j} w_{s_{1},j} \,.\, w_{s_{2},j}}{\sqrt{\sum_{d_j} w^2_{s_{1},j} \cdot \sum_{d_j} w^2_{s_{2},j}}}$
Church and Hanks (1990)	Mutual Information	$log_2\left[\frac{P(s_1, s_2)}{P(s_1)\,.\,P(s_2)} + 1\right]$
Wu and Palmer (1994)	Wu & Palmer similarity	$\frac{2\,.\, d(LCS)}{d(s_1) + d(s_2)}$
Resnik (1995)	Resnik similarity	$-\, log\, p(LCS(s_1, s_2))$
ALMasri et al. (2013)	Semantic similarity	$\frac{\|I(s_1) \cap I(s_2)\| + \|O(s_1) \cap O(s_2)\|}{\|I(s_1) \cup I(s_1)\| + \|O(s_2) \cup O(s_2)\|}$

expansion of term “technology” to “information technology” may work well because “information” is strongly correlated to the overall meaning of the query “data technology”. However, the same reasoning does not apply in the case of the query “music technology”. Bai, Nie, Cao, and Bouchard (2007) discusses the problem of one-to-one association; it deals with query-specific contexts instead of user-centric ones along with the context around and within the query.

In contrast to one-to-one, in the one-to-many association, a candidate term is added to the expanded query if it is correlated to multiple terms from the original query, hence the name “one to many”. Hsu et al. (2006, 2008) use one-to-many association. In these articles, it is compulsory to correlate a new term, extracted from the combination of ConceptNet and WordNet, to a minimum of two original query terms before including the new term into the expanded query. Let q be the original query and let s_2 be an expansion term. In one-to-many association, the correlation coefficient of s_2 with q is calculated as:

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

where:

c_{s_1,s_2} denotes correlation (similarity) score between terms s_1 and s_2, and

$w_{s_1,j}$ ($w_{s_2,j}$) is the weight of the term s_1 (s_2) in the jth document.

Other works based on one-to-many association are Bai et al. (2005); Bhatia et al. (2011); Cui et al. (2003); Gan and Hong (2015); Qiu and Frei (1993); Riezler, Liu, and Vasserman (2008); Sun, Ong, and Chua (2006); Xu and Croft (1996) and Kuzi et al. (2016). As a special mention, Qiu and Frei (1993) and Xu and Croft (1996) have gained large acceptance in literature because of their one-to-many association expansion feature and weighting scheme as described in Eq. (12).

Qiu and Frei (1993) use 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 product of term frequency (tf) of term s in document d_j and the inverse term frequency (itf) of 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 used for document ranking.

Xu and Croft (1996) use concepts (a group of contiguous nouns) instead of individual terms while expanding queries. Concepts are chosen based on term co-occurrence with query terms. Concepts are picked from the top retrieved documents, but they are determined on the basis of the 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-occurrences of query term s_1 in the jth passage and $w_{s_2,j}$ is the frequency of concept s_2 in the jth 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 (Xu & Croft, 1996).

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, as mentioned before, consider queries “data technology” and “music technology”. As discussed before, “information technology” is not an appropriate expansion for the query “music technology”. One way to overcome this problem is by adding *context words* to validate term-term associations. For example, in the case of adding “information” as an expansion term for query “music technology”, an association of “music” and “information” should be considered strong only if these terms co-occur together sufficiently high number of times. Here, “music” is a context word added to evaluate the term-term association of “information” and “technology” in the context of the query “music technology”. Such context words can be extracted from a corpus using term co-occurrence (Bai, Nie, & Cao, 2006; Bai et al., 2007; Jian, Huang, Zhao, He, & Hu, 2016; Wang & Zhai, 2008) or derived from logical significance of a knowledge base (de Boer, Schutte, & Kraaij, 2016; Dalton et al., 2014; Lau, Bruza, & Song, 2004;

Lehmann et al., 2015).

Voorhees (1994) – using WordNet data source for QE – found that the expansion using term co-occurrence techniques is commonly not effective 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 the synset of "incandescent" contains word "light". Thus, instead of the phrase "incandescent light", we can consider the synset of "incandescent". Liu et al. (2004) use this approach for Word Sense Disambiguation (WSD). For example, consider query "tropical storm". In WordNet, the sense of the word "storm" determined through hyponym of the synset {violent storm, storm, tempest} is "hurricane", whose description has the word "tropical". As a result, the sense of "storm" is determined correctly.

For determining the significance of a phrase, Liu et al. (2004) evaluate the correlation value of the terms in the phrase. A phrase will be significant if the terms within the phrase are strongly and positively correlated in the collection of documents. The correlation value of the terms in a phrase $\mathcal{P}$ is described as:

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

where:

$s_1, s_2, \ldots, s_n$ are the terms in $\mathcal{P}$,

$P(\mathcal{P})$ indicates the probability of a document containing $\mathcal{P}$,

$P(s_i)$ is the probability of a document containing an individual term $s_i \in \mathcal{P}$, and

$\prod_{s_i \in \mathcal{P}} P(s_i)$ indicates the probability of a document having all the terms in $\mathcal{P}$ (assuming that these terms are conditionally independent).

For example, consider a collection of 10 documents where the phrase "computer science" is present in only one document. Hence, the probability of a document containing the phrase "computer science" is 0.1. Further, assume that each word in "computer science" also occurs in only one document – in which the phrase occurs. Clearly, such a phrase is very significant when part of a query. The same is confirmed by a high correlation value of 9 computed as per Eq. (13). Liu et al. (2004) suggest the correlation value to be greater than 5 for a phrase to be considered significant.

Another approach for determining one-to-many association is based on the combination of various relationships between term pairs through a Markov chain framework (Collins-Thompson & Callan, 2005). Here, words having the highest probability of relevance in the stationary distribution of the term network are selected for QE. 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). Lv and Zhai (2009) 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. Song, Yu, Wen, and Hon (2011) proposed Proximity Probabilistic Model (PPM) that uses a position-dependent term count to calculate the number of term occurrences and term counts propagated from neighboring terms. Recently, Jian et al. (2016) 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, the first topic model is used for extracting the latent semantic information of the query term. 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) queries.

To overcome the limitations of term-to-term relationships, 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, QE even at phrase level may not offer desired clarity. To discuss this further, consider cases of the phrase being compositional or non-compositional. With the compositional phrases, each and every term associated with the phrase can be expanded using similar alternative terms; the final expanded phrase keeps its significance. Cui et al. (2003) analyzes the phrases using n-grams from the user's query logs. The proposed technique filters the phrases that are not present in the documents being searched. Liu, Li, Zhang, and Xiong (2008) select the most appropriate phrases for QE 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 the query phrase are considered in the set of most appropriate phrases. Recently, Al-Shboul and Myaeng (2014) 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 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. Eguchi (2005) acknowledges that retrieval results are improved when pseudo-relevance feedback is also included in QE based on phrases. The combination of pseudo-relevance feedback and phrase expansion is more effective than phrase expansion alone.

Regarding idiomatic phrases, dealing with them can be troublesome. They are non-compositional in nature and replacing a word with a similar meaning word – as often done during QE – 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", the phrase "break a foot" gives an entirely different meaning from the original phrase.

Table 3 summarizes the mathematical form of term-term correlation value in one-to-many association based on the above discussion.

Table 3
Summary of research work related to one-to-many association QE for term ranking based on term-term correlation values.

Publications	Approaches	Mathematical form
Qiu and Frei (1993), Xu and Croft (1996), Cui et al. (2003), Bai et al. (2005), Sun et al. (2006), Riezler et al. (2008), Bhatia et al. (2011), Gan and Hong (2015), Kuzi et al. (2016)	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)	Correlation value	$\frac{P(\mathcal{P}) - \prod_{s_i\in\mathcal{P}} P(s_i)}{\prod_{s_i\in\mathcal{P}} P(s_i)}$

2.2.3. *Feature distribution of top ranked documents*

Approaches discussed in this section are entirely distinct from the approaches described in earlier sections because the QE techniques discussed 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 QE in response to the initial query. The idea for using the top retrieved documents as a source of potentially relevant documents for a user's domain of interest comes from Attar and Fraenkel (1977). 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 QE 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 QE technique for the modification of the initial query using the terms extracted from the documents in response to the initial query. The user is asked to assess the relevance of the documents retrieved in response to the initial query. The retrieved documents are shown to the user mostly 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 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 QE. The top weighted terms are either added to the initial query automatically or based on manual selection by the user. For example, Java has three synsets with each having a specific sense: island as a geographical place, coffee as a beverage, and as a programming language in computer science. If the query is about Java and the top several retrieved documents are about Java programming, then there may be query drift towards the documents on Java programming. This may not work as desired if the user wants to retrieve documents about Java coffee or Java island. Therefore, if the words added to the original query are unrelated to the query topic, the quality of the retrieval is likely to go down, especially in Web search.

Quite a few term selection techniques have been proposed for QE, which are based on relevance feedback. The common thought behind all these similar techniques is to select terms that will describe the full meaning of the initial query. Rocchio's method (Rocchio, 1971) is one of the first approaches that investigated relevance feedback. This method used an IR 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 the 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,

$\overrightarrow{d_i}$, $\overrightarrow{d_j}$ are relevant and irrelevant document vectors respectively.

In the above paper (i.e., Rocchio, 1971), only the positive feedback documents and their terms were used to modify and expand the initial query; the weights are typically set as $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.15$. Further, any negative term weights are neglected and set to 0.

Jones, Walker, and Robertson (2000) present a probabilistic model for calculating document matching score and came up with superior results on TREC Programme collections. Their approach first 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 the document length normalization factor calculated as $NF = (1 - c) + c \times \frac{DL}{AVDL}$ (c is the tuning constant, DL is the document length, and $AVDL$ is average document length), and,

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).

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 having the query expansion terms and, RW is the relevance weight, which is computed 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 the 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 the relevance feedback approach is Pseudo-relevance feedback (or blind feedback, or retrieval feedback). This directly uses the top retrieved documents in response to the user's initial query for composing query expansion terms. Here, the user is not involved in the selection of relevant documents. Rocchio's method (Rocchio, 1971) 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 the resulting score. The Rocchio'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-relevant document d and

R is the set of pseudo-relevant documents.

However, a disadvantage of the above approach is that it considers the score of each term in the document collection, and in the process, assigns more importance to the whole collection instead 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 the terms having less information content will have nearly the same distribution in any of the documents from the whole collection. Terms that are closely related to the user's query will have a higher probability of occurrence in the retrieved relevant documents.

Various term ranking functions have been proposed on the basis of term distribution in the 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 next.

Robertson and Jones (1976) propose a weighting function known as the 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 in a set of pseudo-relevant documents D_R respectively.

On the same lines, Doszkocs (1978) uses a weighting function known as chi-square (χ^2) for scoring the query terms. It is formulated as:

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

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 in a set of pseudo-relevant documents D_R respectively.

Robertson (1990) 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 the 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:

$w(t,\ d)$ indicates the weight of the term t in pseudo-relevant document d,

R is the set of pseudo-relevant documents, and,

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

On the same lines, Carpineto et al. (2001) use the Kullback–Leibler divergence (KLD) for measuring the term distribution difference between pseudo-relevant documents and the entire document collection. Then, the score of the term is computed by adding the score of the terms having higher scores to the KLD score. 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 $p(t|D_C)$ and $p(t|D_R)$ signify the probability of occurrence of the term t in the document collection D_C and in a set of pseudo-relevant documents D_R respectively.

Using the above term scoring approaches in QE, the experimental studies by Carpineto et al. (2001), Wong, Luk, Leong, Ho, and Lee (2008), and Miao, Huang, and Ye (2012) showed results with marked improvements.

Franzoni and Milani (2012) presented a novel collaborative semantic proximity measurement technique known as PMING distance (further updated in Franzoni, 2017). 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 understand the PMING distance, we first introduce concept similarity measurement techniques: PMI and NGD.

Pointwise Mutual Information (PMI) (Church & Hanks, 1990) 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 an individual event, while MI refers to the average of all possible events. PMI is defined in the same manner 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) Cilibrasi and Vitanyi (2007) 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 to any other search engine. NGD between 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 $f(s_1,\ s_2)$ denote the number of results returned by the 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 well as a proximity measurement technique, in a strict sense it cannot be considered as a metric because it does not satisfy the property of triangular inequality.

PMING distance (Franzoni, 2017; Franzoni & Milani, 2012) includes the combination of two semantic similarity measurement techniques: PMI and NGD. PMING distance is defined as a convex linear combination of locally normalized PMI and NGD 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 $PMING\colon W \times W \rightarrow \lceil 0,\ 1 \rceil$ 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 ≤ ρ ≤ 1,

N is the total number (if known) or estimated number (if unknown) of documents indexed by the search engine,

μ_1 and μ_2 are constants; their values depend on the context of evaluation W, and are defined as:

$$\mu_1 = \max_{s_1,s_2 \in W} PMI_{s_1,s_2} \tag{26}$$

$$\mu_2 = \max_{s_1,s_2 \in W} NGD_{s_1,s_2} \tag{27}$$

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, Paik, Pal, and Parui (2014) 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 is defined as:

$$Score(t, F^q) = log_2(df(t, F^q)) \times idf(t, C) \tag{28}$$

where:

F^q is the set of feedback documents for the query q,

$df(t, F^q)$ indicates the number of documents in F^q having the term t, and,

idf stands for inverse document frequency, which is defined as $idf(t, C) = log\frac{N}{df(t, C)}$ (N is the number of documents in the whole collection and $df(t, C)$ corresponds to the document frequency of the term t in the collection C).

Every term ranking method has its own motivation, and the outcomes offered by their utilization are also distinct. In the case of specific queries, it has been observed that the organized sets of expansion terms recommended for each query are mostly unrelated to the original query (Carpineto, Romano, & Giannini, 2002). However, various experimental analyses (such as by Harman, 1992, Salton & Buckley, 1997, Carpineto et al., 2001, and Miao et al., 2012) observe that the selection of the ranking approach commonly does not have an enormous significance on the system efficiency; it is just an approach to determine the set of terms for QE.

2.2.4. Query language modeling

In this approach for QE, a statistical language model is constructed that assigns a probability distribution over the term-collections. Terms with maximum probability are chosen for QE. This approach is also known as the model-based approach. The two popular foundation language models are relevance model (based on the probabilities of the terms in the relevant documents) (Croft & Lafferty, 2013; Lavrenko & Croft, 2001) and mixture model (Zhai & Lafferty, 2001); both are based on the top retrieved documents for QE.

In the relevance-based language model, Lavrenko and Croft (2001) has caught the attention of researchers with their robust probabilistic approach. Their approach assumes that a query q_i and its 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 the conditional probability of the initial query term q_i ($i \in 1, \ldots, n$). The probability of term t in the relevant documents is computed as:

$$p(t|M_{rel}) = \sum_{\theta_d \in R} p(\theta_d) p(t|\theta_d) \prod_{i=1}^{n} p(q_i|\theta_d) \tag{29}$$

In the above Eq. (29), it is 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 QE. This model does not depend upon the distribution difference analysis; hence, it can be said that conceptually this model is very much like Rocchio's method (Rocchio, 1971). The main difference of this model from the Rocchio's is that the top retrieved documents are assigned a weight such that the lower ranked documents have an insignificant impact on the term probability (Lavrenko & Allan, 2006).

In the research area of relevance model, several studies have been published (Bendersky, Metzler, & Croft, 2011; Lv & Zhai, 2010; Miao et al., 2012). Lv and Zhai (2009) performed a correlative analysis on several states of pseudo-relevance feedback and concluded that the relevance model is the most efficient method for the selection of expansion terms. Bendersky et al. (2011) use external resources for generating features for weighting different types of query concepts and consider the latent concepts for expanding the initial query. Miao et al. (2012) proposed a proximity-based feedback model that is based on the traditional Rocchio's model, known as PRoc. It focuses on the proximity of terms rather than the positional information (unlike position relevance model (PRM) (Miao et al., 2012)). It calculates the weights of the candidate expansion terms by taking their distance from the query terms into account. Metzler and Croft (2007) consider term dependencies during QE; the expansion technique is based on Markov random fields model. This model provides a robust framework that includes both term occurrence and proximity-based features. An example of a Markov random field is estimating the number of times the query terms occur within a window of fixed size in an organized or unorganized way. Lv and Zhai (2010) present 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 the query terms in the feedback documents. As another step in improving the research on relevance model, Dalton and Dietz (2013) present a neighborhood relevance model that uses relevance feedback approaches for recognizing the specialty of entity linking across the document and query collections. Actually, the primary objectives of entity linking are to map a string in a document to its entity in knowledge base and to recognize the disambiguating context inside the knowledge base. Recently, Dang, Luk, and Allan (2016) proposed a context-dependent relevance model that provides an approach to incorporate the feedback through improvement of document language models. For evaluating document language models, it uses the context information on the relevant or irrelevant documents to obtain the weight counts (using BM25 weights (Robertson, 2004; Robertson & Walker, 1994)) of the individual query terms.

Discussing the mixture model method, Zhai and Lafferty (2001) consider 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

$$log\, p(T_R|\theta_Q) = \sum_{D\in T_R}\sum_t c(t,\, D)\log(\lambda\, p(t|C) + (1-\lambda)p(t,\,\theta_Q)) \tag{30}$$

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, Expectation Maximization algorithm (Dempster, Laird, & Rubin, 1977) 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 the inadequacy of retrieved documents. Comparatively, this mixture model has a stronger theoretical justification, however, estimating the value of weighting parameter λ can be a difficult task.

Comparative analysis: Among the all weighing and ranking techniques discussed earlier, one-to-many association technique has been used widely. However, weighting and ranking techniques depended upon the different characteristics of the query terms and the data sources used. 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, this may not accurately demonstrate the relationship between an expansion term and the query as a whole. For example, "break a leg" is a theatrical slang meaning "good luck!". When we replace "leg" with synonym "foot", the phrase "break a foot" gives an entirely different meaning from the original phrase. For resolving such language ambiguity problem, one-to-many association plays a crucial role. It correlates the entire query or considers multiple terms from the user's query by assigning correlation score. However, for assigning the weight to the individual terms, one-to-one association play a crucial role. In one-to-one association weighting technique, Jaccard coefficient (Jaccard, 1912) and Cosine similarity (Attar & Fraenkel, 1977) are used widely for assigning the weight to the expansion terms. The weighting techniques discussed under the category of *feature distribution of top-ranked documents* are entirely distinct from the rest because they consider the expansion terms that chosen from the top retrieved documents. However, a disadvantage of the above approach is that it considers the score of each term in the document collection, and in the process, assigns more importance to the whole collection instead of the user's query. Among all the weighting techniques in this category, Rocchio (1971), Robertson Selection Value (RSV) (Robertson & Jones, 1976) and Kullback–Leibler Divergence (KLD) (Carpineto et al., 2001) weighting techniques have been used widely for weighting the expansion terms. Query language modeling is a probabilistic weighting technique that assigns a probability distribution over term-collections. Terms with maximum probability are chosen for QE. Recently, this technique has been used widely in QE. This model does not depend upon the distribution difference analysis; hence, it can be said that conceptually this model is very much like Rocchio's method (Rocchio, 1971).

Table 4 summarizes some important term similarity scores in mathematical form for term ranking based on the above discussion.

2.3. Selection of query expansion terms

In the previous Section 2.2, weighting and ranking of expansion terms were discussed. After this step, the top-ranked terms are selected for QE. The term selection is done on an individual basis; mutual dependence of terms is not considered. This may be debatable; however, some experimental studies (e.g, Lin & Murray, 2005) suggest that the independence assumption may be empirically equitable.

Table 4
Summery of approaches for term ranking based on the term similarity score.

Reference	Approach	Mathematical form
Rocchio (1971)	Rocchio's weights	$\Sigma_{d\in R}w(t,\, d)$
Robertson and Jones (1976)	Dice coefficient	$log\frac{p(t\mid D_R)[1-p(t\mid D_C)]}{p(t\mid D_C)[1-p(t\mid D_R)]}$
Doszkocs (1978)	Chi-square (χ^2)	$log\frac{[p(t\mid D_R)-p(t\mid D_C)]^2}{p(t\mid D_C)}$
Robertson (1990)	Robertson selection value (RSV)	$\sum_{d\in R} w(t,\, d).\ [p(t\vert D_R) - p(t\vert D_C)]$
Carpineto et al. (2001)	Kullback–Leibler divergence (KLD)	$p(t\vert D_R).\ log\frac{p(t\mid D_R)}{p(t\mid D_C)}$
Zhai and Lafferty (2001)	Log-likelihood	$log\, p(T_R\vert\theta_Q) = \sum_{D\in T_R}\sum_t c(t,\, D)\, log(\lambda\, p(t\vert C) + (1-\lambda)p(t,\,\theta_Q))$
Cilibrasi and Vitanyi (2007)	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), Franzoni (2017)	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)	Scoring function	$Score(t,\, F^q) = log_2(df(t,\, F^q)) \times idf(t,\, C)$

Table 5
Summery of terms selection suggested by several researchers.

Number of terms	Reference
One third of the terms	Robertson and Willett (1993)
5–10 terms	Amati et al. (2003), Chang et al. (2006)
20–30 terms	Harman (1992), Zhang et al. (2016)
30–40 terms	Paik et al. (2014)
Few hundreds terms	Bernardini and Carpineto (2008), Wong et al. (2008)
350–530 terms	Buckley et al. (1995)

It may happen that the chosen QE technique produces a large number of expansion terms, but it might not be realistic to use all of these expansion terms. Usually, only a limited number of expansion terms are selected for QE. This is because the IR effectiveness of a query with a small set of expansion terms is usually better than the query having a large set of expansion terms (Salton & Buckley, 1997); this happens due to noise reduction.

While researchers agree that the addition of selective terms improves the retrieval effectiveness, however, the suggested optimum number can vary from a few terms to a few hundred terms. There are different point of views on the number of selective terms to be added: one-third of the original query terms (Robertson & Willett, 1993), five to ten terms (Amati, Joost, & Rijsbergen, 2003; Chang, Ounis, & Kim, 2006), 20–30 terms (Harman, 1992; Zhang et al., 2016), 30–40 terms (Paik et al., 2014), few hundreds of terms (Bernardini & Carpineto, 2008; Wong et al., 2008), 350–530 terms for each query (Buckley, Salton, Allan, & Singhal, 1995). The source of these terms can be the top retrieved documents or well known relevant documents. It has been found that the addition of these expansion terms improves retrieval effectiveness by 7% to 25% (Buckley et al., 1995). On the contrary, some studies show that the number of terms used for QE is a less important factor than the terms selected on the basis of types and quality (Sihvonen & Vakkari, 2004). It has been commonly shown that the effectiveness of QE decreases minutely with the number of non-optimum expansion terms (Carpineto et al., 2001). Most of the experimental studies observed that the number of expansion terms is of low relevance and it varies from query to query (Billerbeck et al., 2003; Billerbeck & Zobel, 2004; Buckley & Harman, 2004; Cao, Nie, Gao, & Robertson, 2008). It has been observed that the effectiveness of QE (measured as mean average precision) decreases when we consider less than 20 expansion terms (Paik et al., 2014; Zhang et al., 2016). Usually, 20–40 terms are the best choice for QE. Zhai and Lafferty (2001) assign a probability score to each expansion term and select those with a score higher than a fixed threshold value $p = 0.001$ (Table 5).

However, instead of considering an optimum number of expansion terms, it may be better to adopt more informed selection techniques. Focus on the selection of the most relevant terms for QE instead of an optimum number of terms yields better results (Cao, Nie et al., 2008; Carpineto et al., 2002).

For the selection of the expansion terms on the basis of the ranks assigned to the individual term, various approaches have been proposed that exploit the additional information. Carpineto et al. (2002) proposed a technique that uses multiple term ranking functions and selects the most common terms for each query. A similar approach is utilized by Collins-Thompson and Callan (2007); 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, which are related to the various query features. Another approach for selecting expansion terms that depend upon the query ambiguity has been proposed by Chirita, Firan, and Nejdl (2007). Here, the number of expansion terms depends on the ambiguity of the initial query in the web or the user log; the ambiguity is determined by the clarity score Cronen-Townsend, Zhou, and Croft (2002). Cao, Nie et al. (2008) use a classifier to recognize relevant and irrelevant expansion terms. Whether the classifier parameter works well or not for labeling the individual expansion terms, depends on the effectiveness of the retrieval performance and the co-occurrence of the query terms. Their study shows that the top retrieved documents contain as many as 65% harmful terms. For selecting the best expansion terms, Collins-Thompson (2009) optimized the retrieved data with respect to uncertainty sets resulting in an optimization problem.

In spite of the above, it has been shown that the majority of existing works on QE (Lavrenko & Allan, 2006; Wu & Fang, 2013) only focus on indexing and document optimization for the selection of expansion terms, and neglect the re-ranking score. However, recently a number of articles (Diaz, 2015; Zhang et al., 2016) supported the re-ranking with valid proof and obtained good retrieval effectiveness. Wu and Fang (2013) proposed impact-sorted indexing technique that utilizes a particular index data structure; the technique improves the scoring methods in IR. Lavrenko and Allan (2006) use the pre-calculated pairwise document similarities to reduce the searching time for the expanded queries. However, supporting re-ranking, Diaz (2015) points out that re-ranking can provide nearly identical performance as the results returned from the second retrieval done using the expanded query. This works specifically for precision-oriented metrics. This has also been verified in experimental results of Zhang et al. (2016), who utilize re-ranking as the default approach for IR. They also suggests to add 20 to 30 expansion terms in the initial query to improve the IR effectively.

2.4. Query reformulation

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

A popular query reweighting method was proposed by Salton and Buckley (1990), which is influenced by Rocchio's method (Rocchio, 1971) for relevance feedback and its consequential developments. It can be formulated as:

$$w'_{t,q_e} = (1 - \lambda) .\ w_{t,q} + \lambda .\ W_t \tag{31}$$

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 the weighting parameter that balances the comparative contribution of the original query terms (q) and the expansion terms.

When Rocchio's weights (see Eq. (18)) are used for calculating the weights of the QE terms that are extracted from the pseudo-relevant documents, it can be observed that the expanded query vector measured by Eq. (31) is relevant to the pseudo-relevant documents. This reduces the term distribution difference between the pseudo-relevant documents and the documents having expansion terms reweighted by Rocchio's weighting scheme. The intention is to assign a 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 such as Asian (Savoy, 2005), European (Amati et al., 2003; Darwish, Magdy et al., 2014; Larkey, Ballesteros, & Connell, 2007), Hindi (Bhattacharya, Goyal, & Sarkar, 2016; Paik et al., 2014), and other languages (Carpineto et al., 2001; Paik et al., 2014; Wong et al., 2008; Zhang et al., 2016). It has been observed that the reweighting system based on inverse term ranks also provides a favorable outcome (Carpineto et al., 2002; Hu, Deng, & Guo, 2006). 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 Eq. (31), their values must be normalized. A number of normalization approaches have been discussed in the survey by Wong et al. (2008); it was observed that the discussed approaches commonly provide similar outcomes. However, Montague and Aslam (2001) observe the need for a better approach that normalizes not only data but also increases equality among normalized terms, which can be more expressive.

In addition to the above discussion, the value of the weighting parameter (λ) in Eq. (31) should be adjusted appropriately for improving retrieval effectiveness. A common choice is to grant more significance (e.g., multiply by two) 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 by Amati et al. (2003). Another effective approach is to compute the weights, to be assigned to the expansion terms, query-wise. For example, Lv and Zhai (2009, 2010) use relevance feedback in combination with a learning approach for forecasting 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 only with the feedback documents. However, Eq. (31) 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 the path length, number of co-occurrences, number of connections and relationship types (Jones et al., 1995). For example, Voorhees (1994) 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, Hsu et al. (2008) use activation score for weighting of expanded terms. The weight of an expansion term is often decided by its correlation or similarity with the considered query.

When document ranking is based on the language modeling approach (see Section 2.2.4), the query reweighting step usually favorably expands the original query. In the language modeling framework, the most relevant documents are the ones that decrease Kullback–Leibler divergence (KLD) between the document language model and the query language model. It is formulated as:

$$Sim_{KLD}(Q, D) \propto \sum_{t \in V} p(t|\theta_Q).\ \log \frac{p(t|\theta_Q)}{p(t|\theta_D)} \tag{32}$$

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 (Jelinek, 1980; Jelinek & Mercer, 1980):

$$p(t|\theta'_D) = (1 - \lambda) .\ p(t|\theta_D) + \lambda .\ p(t|\theta_C) \tag{33}$$

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.

Eq. (32)raises the following question: is it possible to build a better query model by obtaining similar terms using their concern-probabilities? Further, will it smoothen the original query model using the equivalent expanded query model (EQM) just as collection model θ_C smoothens the document model based on Eq. (33)? To answer this, several approaches have been proposed to build an expanded query model that not only considers feedback documents (Lavrenko & Croft, 2001; Zhai & Lafferty, 2001) but also term relations (Bai et al., 2005; Cao, Gao, Nie, & Bai, 2007; Gan & Hong, 2015) and domain hierarchies (Bai et al., 2007), and can be heuristic (Shah & Croft, 2004). Hence, Carpineto and Romano (2012) suggested that instead of considering a particular method, one can come up with a superior expanded query model (calculated using Jelinek–Mercer interpolation Jelinek & Mercer, 1980) given as:

$$p(t|\theta'_Q) = (1-\lambda)\,.\; p(t|\theta_Q) + \lambda\,.\; p(t|\theta_{EQM}) \tag{34}$$

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

This equation is the probabilistic representation of Eq. (31) and many articles (Carpineto & Romano, 2012; Gan & Hong, 2015; Kotov & Zhai, 2012; Kuzi et al., 2016; Xu et al., 2009; Zhang et al., 2016) have used it for probabilistic query reweighting.

Though the query reweighting approach is generally used in QE 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 (Carpineto & Romano, 2012). Another way can be first to increase the number of similar query terms, and then apply a customized weighting function for ranking the expanded query terms; instead of using the fundamental weighting function used for reweighting the expanded query. This technique was used by Robertson and Walker (Robertson & Walker, 2000) to enhance the Okapi BM25 ranking function (Robertson, Walker, Beaulieu, & Willett, 1999). Some other approaches for query reformulation are utilization of structured query (Collins-Thompson & Callan, 2005; Jamil & Jagadish, 2015; Kato, Sakai, & Tanaka, 2012; Pound, Ilyas, & Weddell, 2010), Boolean query (Graupmann, Cai, & Schenkel, 2005; Kim, Seo, & Croft, 2011; Liu et al., 2004; Pane & Myers, 2000), XML query (Chu-Carroll, Prager, Czuba, Ferrucci, & Duboue, 2006; Junedi, Genevès, & Layaïda, 2012; Kamps, Marx, Rijke, & Sigurbjörnsson, 2006) and phrase matching (Arguello et al., 2008).

3. Importance and application of query expansion

3.1. Importance of query expansion

One of the major importance of QE is that it enhances the chance to retrieve the relevant information on the Internet, which is not retrieved otherwise using the original query. Many times the 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 Internet searches. For example, if the user's original query is "*Novel*", it is not clear what the user wants: the user may be searching for a fictitious narrative book, or the user may be interested in something new or unusual. Here, QE can expand the original query *"Novel"* to {*"Novel book", "Book"*}, or to {*"New", "Novel approach"*} depending upon the user's interest. The new queries retrieve documents specific to the two types of meaning. This technique has been used hugely for search operations in various commercial domains (e.g., education, hospitality in medical science, economics and experimental research (Carpineto & Romano, 2012)), where the primary goal is to retrieve all documents relevant to a particular concern.

The above fact that the use of QE to retrieve a lot of relevant documents increases recall rate, it adversely affects precision is also well supported experimentally (Harman & Voorhees, 1996; He & Ounis, 2009; Pakhomov, Finley, McEwan, Wang, & Melton, 2016). 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 the ranking after QE. For improvement of retrieval precision, expanded query can also use Boolean operators (AND,OR) or PARAGRAPH operator (Moldovan & Mihalcea, 2000) to transform the expanded query to Boolean query (Kim et al., 2011; Pane & Myers, 2000), 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 $t'_m\}$, $B_{query}= \{t_1$ 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 the recall rate, whereas, OR operator reduces precision but improves the recall rate (Turtle, 1994). Kim et al. (2011) propose a novel Boolean query suggestion technique where Boolean queries are produced by exploiting decision trees learned from pseudo-labeled documents. The produced queries are ranked using query quality predictors. The authors compared this technique to contemporary QE techniques and experimentally demonstrated the formers' superiority. XML queries can also be used for improving precision in IR systems for enhancing the Internet searches (e.g., Chu-Carroll et al., 2006; Junedi et al., 2012; Kamps et al., 2006). Improving precision in IR systems through QE using web pages has been proposed by Cui et al. (2002, 2003) and Zhou, Lawless, and Wade (2012). Here, QE happens based on a collection of important words in related web pages. Another set of techniques for the same task expand queries using query concept (Dalton et al., 2014; Fonseca, Golgher, Pôssas, Ribeiro-Neto, & Ziviani, 2005; Hsu et al., 2008; Qiu & Frei, 1993). Here, expansion happens based on the similar meaning of query terms.

However, when considering the joint evaluation of the precision and recall rates in QE, several experimental studies agree that the expansion of the user query enhances the average precision of the query results by ten percent or more (e.g., Büttcher, Clarke, & Cormack, 2016; Carpineto et al., 2002; Collins-Thompson, Macdonald, Bennett, Diaz, & Voorhees, 2015; Egozi, Markovitch, & Gabrilovich, 2011; Lee et al., 2008; Rivas, Iglesias, & Borrajo, 2014; Salton & Buckley, 1997; Xu & Croft, 2000). These experimental studies also support the effectiveness of QE techniques in IR systems. Some recent studies have shown that QE can also improve the precision by disambiguating the user query (e.g., Bai et al., 2005; Stokoe, Oakes, & Tait, 2003; Yao, Yi, Liu, Zhao, & Sun, 2015; Zhong & Ng, 2012). Table 6 summarizes some prominent works in literature towards improving the precision and recall rates.

3.2. Application of query expansion

Beyond the key area of IR, there are other recent applications where QE techniques have proved beneficial. We discuss some of such applications next.

Table 6
Summary of techniques used for improving precision & recall rate.

Expansion Techniques	Publications
Boolean query	Pane and Myers (2000), Moldovan and Mihalcea (2000), Kim et al. (2011)
XML queries	Kamps et al. (2006), Chu-Carroll et al. (2006), Junedi et al. (2012)
Collection of the top terms within the web pages	Cui et al. (2002), Cui et al. (2003), Zhou et al. (2012)
Query concepts	Qiu and Frei (1993), Fonseca et al. (2005), Hsu et al. (2008), Bouchoucha et al. (2013)
Query disambiguation	Stokoe et al. (2003), Bai et al. (2005), Zhong and Ng (2012), Yao et al. (2015)

3.2.1. *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 his interests and commitments. This has generated a need of a QE framework that is based on social bookmarking and tagging systems, which enhance the document representation.

Bender et al. (2008) present a QE framework to exploit the different entities present in social relations (users, documents, tags) and their mutual relationships. It also derives scoring functions for each of the entities and their relations. Biancalana and Micarelli (2009) use social tagging and bookmarking in QE for personalized web searches. Their experimental results show effective matching of the user's interests with the search results. Bouadjenek, Hacid, Bouzeghoub, and Daigremont (2011) use a combination of social proximity and semantic similarity for personalized social QE, which is based on similar terms that are mostly used by a given user and his social relatives. Zhou et al. (2012) proposed a QE technique that is based on distinctive user profiles in which the expansion terms are extracted from both the annotations and the resources that the user has created and opted (also used by Bouadjenek, Hacid, & Bouzeghoub, 2013a). Many other works (e.g., Bouadjenek, Hacid, & Bouzeghoub, 2013b; Hahm, Yi, Lee, & Suh, 2014; Mulhem, Amer, & Géry, 2016) discuss the QE and social personalized ranking in the context of personalized social documents. Recently, Bouadjenek et al. (2016) proposed a technique PerSaDoR (**Per**sonalized **S**oci**a**l **Do**cument **R**epresentation) that uses (i) the user's activities in a social tagging system for indexing and modeling, and, (ii) social annotations for QE. A more recent work in personalized IR by Amer, Mulhem, and Géry (2016) uses word embedding for QE. Here, the experimental evaluation was done on the collection of CLEF Social Book Search 2016.[8] The main motive of this paper is to address the following questions: (1) "How to use the word embedding technique for QE in the context of the social collection?" and (2) "How to use the word embedding technique to personalize QE?". Zhou, Wu, Zhao, Lawless, and Liu (2017) personalized QE using enriched user profiles on the web; the user profiles were created using external corpus "folksonomy data". They also proposed a model to enhance the user profiles. This model integrates word embeddings with topic models in two groups of pseudo-relevant documents: user annotations and documents from the external corpus.

3.2.2. *Question answering*

Question answering (QA) has become a very influential research area in the field of IR systems. The primary objective of QA is to grant a quick answer in response to the user's query. Here, the focus is to keep the answer concise rather than retrieving all relevant documents. As input, the system accepts questions (instead of a set of terms) in natural language, e.g., "Which is the first nation in the world to enter Mars orbit in the first attempt?". Recently, search engines have also started using the QA system to provide answers to such types of questions. However, for ranking the answers of such questions, the main challenges in QE is the mismatch problem, which arises due to a mismatch between the expression in question and the text-based answers (Lin & Pantel, 2001).

To overcome the mismatch problem and to improve the document retrieval in QA systems, a lot of research has been done. In 2004, Agichtein, Lawrence, and Gravano (2004) presented an important approach for QE using FAQ data. The purposed system automatically learns to transform natural language questions into queries with the goal of maximizing the probability of an information retrieval system returning documents that contain answers to a given question; the same approach was also followed by Soricut and Brill (2006). Riezler, Vasserman, Tsochantaridis, Mittal, and Liu (2007) present a technique to expand the user's original query in QA system using Statistical Machine Translation (SMT) to bridge the lexical gap between questions and answers. SMT attempts to link the linguistic difference between the user's query and system's response. The goal of this system is to learn lexical correlations between words and phrases in questions and answers. Bian, Liu, Agichtein, and Zha (2008) present a ranking framework to take advantage of user interaction information to retrieve high-quality and relevant content in social media. It has ranked the retrieved documents in QA using community-based features and, user preference of social media search and web search. Other works (Cavalin et al., 2016; Liu, Chen, Shen, & Lu, 2015; Molino, Aiello, & Lops, 2016; Panovich, Miller, & Karger, 2012) use social network for improving the retrieval performance in QA. Wu et al. (2014) expand short queries by mining the user intent from three different sources, namely community question answering (CQA) archive, query logs, and web search results. Currently, QE in **Q**uestion **A**nswering over **L**inked open **D**ata (QALD) has gained much attention in the area of natural language processing. Shekarpour, Höffner, Lehmann, and Auer (2013) 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)[9] cloud. The evaluation was carried out on a training dataset extracted from the QALD[10] question answering benchmark dataset. The

[8] http://social-book-search.humanities.uva.nl .
[9] http://lod-cloud.net/ .
[10] http://qald.sebastianwalter.org/ .

experimental results show a considerable improvement in precision and recall rates over the baseline approaches.

3.2.3. *Cross-Language Information Retrieval (CLIR)*

It is a part of IR that retrieves the information present in a language(s) different from the user's query-language. For example, a user can query in Hindi, but the retrieved relevant information can be in English. Over the past few years, CLIR has received significant attention due to the popularity of CLEF[11] 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 (Radwan, 1994), parallel or comparable corpora-based techniques (Sheridan & Ballerini, 1996) and machine-readable bilingual dictionaries (Ballesteros & Croft, 1996). Research challenges with the traditional CLIR are untranslatable query terms, phrase translation, inflected term and uncertainty in language translation between the source and target languages (Pirkola, Hedlund, Keskustalo, & Järvelin, 2001). To overcome this translation error, a popular approach is to use QE (Ballesteros & Croft, 1997; Nie, Simard, Isabelle, & Durand, 1999). It gives a better output – even in the case of no translation error – due to the use of statistical semantic similarity among the terms (Adriani & Van Rijsbergen, 1999; Kraaij, Nie, & Simard, 2003). To counter the errors in the automated machine translation in the case of cross-language queries, Gaillard, Bouraoui, De Neef, and Boualem (2010) use linguistic resources for QE. QE can be applied at various points in the translation process: before or after translation, or both. It has been shown that the application at prior translation gives better result in comparison to the application at post-translation, however, the application at either step gives superior results in comparison to not using QE (Ballesteros & Croft, 1998; Ballesteros, 2002; Levow, Oard, & Resnik, 2005; McNamee & Mayfield, 2002). For improving the QE in CLIR, Cao et al. (2007) combines the dictionary translation and the co-occurrence term-relations into Markov Chain (MC) models, which is defined as a directed graph where query translation is formulated as a random walk in the MC models. Recently, Zhou, Lawless, Liu, Zhang, and Xu (2015); Zhou et al. (2016) used QE techniques to personalize CLIR based on user's historical usage information. Experimental results show that personalized approaches work better than non-personalized approaches in CLIR. Bhattacharya et al. (2016) present a technique to translate user queries from Hindi to English CLIR using word embedding. The proposed word embedding based approach captures contextual clues for a particular word in the source language and gives those words as translations that occur in a similar context in the target language.

3.2.4. *Information filtering*

Information filtering (IF) is a method to eliminate non-essential 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, etc. (for a survey, see Hanani, Shapira, & Shoval, 2001). There are two basic approaches for IF: content-based filtering and collaborative filtering. Belkin and Croft (1992) discusses the relationship between IR and IF, and, establish that IF is a special kind of IR with the same set of research challenges and outcomes. They have a common goal to provide relevant information to the user but from different aspects. Hanani et al. (2001) give 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 expand the user's query in a way that can well reflect the user's interest and needs (Allan, 1996). Eichstaedt et al. Eichstaedt, Patel, Lu, Manber, and Rudkin (2002) combine user's query with system's master query for improving results. Other techniques include using user's profile (Yu, Tresp, & Yu, 2004), using geographical query footprint (Fu, Jones, & Abdelmoty, 2005), using correlated keywords (Zimmer, Tryfonopoulos, & Weikum, 2008), using links and anchor texts in Wikipedia (Arguello et al., 2008), using text classification in twitter messages (Sriram, Fuhry, Demir, Ferhatosmanoglu, & Demirbas, 2010) and using the behavior-patterns of online users (Gao, Xu, & Li, 2015; Zhang & Zeng, 2012). Recently, Wu, Liu, Xie, Ester, and Yang (2016) reformulated the query using the user-item co-clustering method for improving the Collaborative Filtering technique. Another work by Zervakis, Tryfonopoulos, Skiadopoulos, and Koubarakis (2017) reorganizes query using DBpedia[12] and ClueWeb09[13] corpora for efficient Boolean IF. The proposed approach uses linguistically motivated concepts, such as words, to support continuous queries that are comprised of conjunctions of keywords. These continuous queries may be used as a basis for query languages that support not only basic Boolean operators but also more complex constructs, such as proximity operators and attributes.

3.2.5. *Multimedia Information Retrieval*

Multimedia Information Retrieval (MIR) deals with searching and extracting the semantic information from multimedia documents, such as audio, video and image (Lew, Sebe, Djeraba, & Jain, 2006 gives a good survey in this regard). For IR in multimedia documents, most of the MIR systems typically rely on text-based searches, 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 precisely describe the actual multimedia content. Hence, QE 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 in 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 a text search on the transcription of the audio file. However, the transcription is obtained automatically by speech translation software, and hence, contains errors. In such a case, expanding the transcription by adding related words greatly improves the retrieval effectiveness (Singhal &

[11] http://www.clef-initiative.eu/ .

[12] http://wiki.dbpedia.org/ .

[13] http://lemurproject.org/clueweb09/ .

Pereira, 1999). However, for text document retrieval, the benefits of such a document expansion are limited (Wei & Croft, 2007). Jourlin, Johnson, Jones, and Woodland (1999) show that QE can improve the average precision by 17 percent in audio retrieval. Barrington, Chan, Turnbull, and Lanckriet (2007) follow QE technique based on semantic similarity in audio IR. Tejedor et al. (2012) compare language dependent and language independent queries through examples and conclude that the language dependent setup provides better results in spoken term detection. Recently, Khwileh and Jones (2016) presented a QE method for social audio contents, where the QE 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 the manually established text descriptions of the visual concepts. Natsev, Haubold, Tešić, Xie, and Yan (2007) expand text query using lexical, statistical and content-based approaches for visual QE. Tellex, Kollar, Shaw, Roy, and Roy (2010) expand queries using the corpus of natural language description based on the accurate evaluation of system performance. A more recent work (Thomas, Gupta, & Venkatesh, 2016) uses 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 the image descriptions in the database (for reviews on image retrieval see Li et al., 2016 and Datta, Joshi, Li, & Wang, 2008). Kuo, Chen, Chiang, and Hsu (2009) present two QE approaches: intra expansion (expanded query is obtained from existing query features) and inter expansion (expanded query is obtained from the search results). Hua et al. (2013) use the query logs data for generic web image searches. Borth, Ji, Chen, Breuel, and Chang (2013) use multitag for image retrieval, whereas, Liu, Yan, Ji, Hua, and Zhang (2013) retrieve images using a query adaptive hashing method. Xie, Zhang, Tan, Guo, and Li (2014) present a contextual QE technique to overcome the semantic gap of visual vocabulary quantization, and, performance and storage loss due to QE in image retrieval.

3.2.6. *Other applications*

Some other recent applications of QE are plagiarism detection (Nawab, Stevenson, & Clough, 2016), event search (Atefeh & Khreich, 2015; Boer, Schutte, & Kraaij, 2015; Douze, Revaud, Schmid, & Jégou, 2013), text classification (Wang et al., 2016), patent retrieval (Magdy & Jones, 2011; Mahdabi & Crestani, 2014; Wang & Lin, 2016), dynamic process in IoT (Huber, Seiger, Kühnert, Theodorou, & Schlegel, 2016; Huber, Seiger, Kühnert, & Schlegel, 2016), classification of e-commerce (Jammalamadaka, Salaka, Johnson, & King, 2015), biomedical IR (Abdulla, Lin, Xu, & Banbhrani, 2016), enterprise search (Liu, Chen, Fang, & Wang, 2014), code search (Nie, Jiang, Ren, Sun, & Li, 2016), parallel computing in IR (Macfarlane, Robertson, & Mccann, 1997) and twitter search (Kumar & Carterette, 2013; Zingla, Chiraz, & Slimani, 2016).

Table 7summarizes some of the prominent and recent applications of QE in literature based on the above discussion.

Table 7
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), Bouadjenek et al. (2016), Amer et al. (2016), Mulhem et al. (2016), Hahm et al. (2014), Bouadjenek et al. (2013a), Zhou et al. (2012), Bouadjenek et al. (2011), Biancalana and Micarelli (2009)
Question answering	FAQs, QA pairs, Social network, WordNet, LOD cloud, community question answering (CQA) archive, query logs and web search	Respond to user's query with quick concise answers rather than retunring all relevant documents	Molino et al. (2016), Cavalin et al. (2016), Liu et al. (2015), Shekarpour et al. (2013), Panovich et al. (2012), Bian et al. (2008), Riezler et al. (2007)
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), Bhattacharya et al. (2016), Zhou et al. (2015), Gaillard et al. (2010), Cao et al. (2007), Kraaij et al. (2003)
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. (2017), Wu et al. (2016), Gao et al. (2015), Zhang and Zeng (2012), Arguello et al. (2008), Fu et al. (2005), Yu et al. (2004)
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), Thomas et al. (2016), Li et al. (2016), Xie et al. (2014), Liu et al. (2013), Tejedor et al. (2012), Tellex et al. (2010), Kuo et al. (2009), Wei and Croft (2007)
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), Wang and Lin (2016), Nawab et al. (2016), Huber, Seiger, Kühnert, Theodorou, and Schlegel (2016), Zingla et al. (2016), Abdulla et al. (2016), Nie et al. (2016), Atefeh and Khreich (2015), Liu et al. (2014)

4. Classification of query expansion approaches

On the basis of data sources used in QE, 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 analyses can be further split into four and two subclasses respectively as shown in Fig. 5. This section discusses the QE approaches based on the properties of various data sources used in QE as shown in Fig. 5.

4.1. Global analysis

In the global analysis, QE techniques implicitly select expansion terms from hand-built knowledge resources or from large corpora for reformulating the initial query. Only individual query terms are considered for expanding the initial query. The expansion terms are semantically similar to the original terms. Each term is assigned a weight; the 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

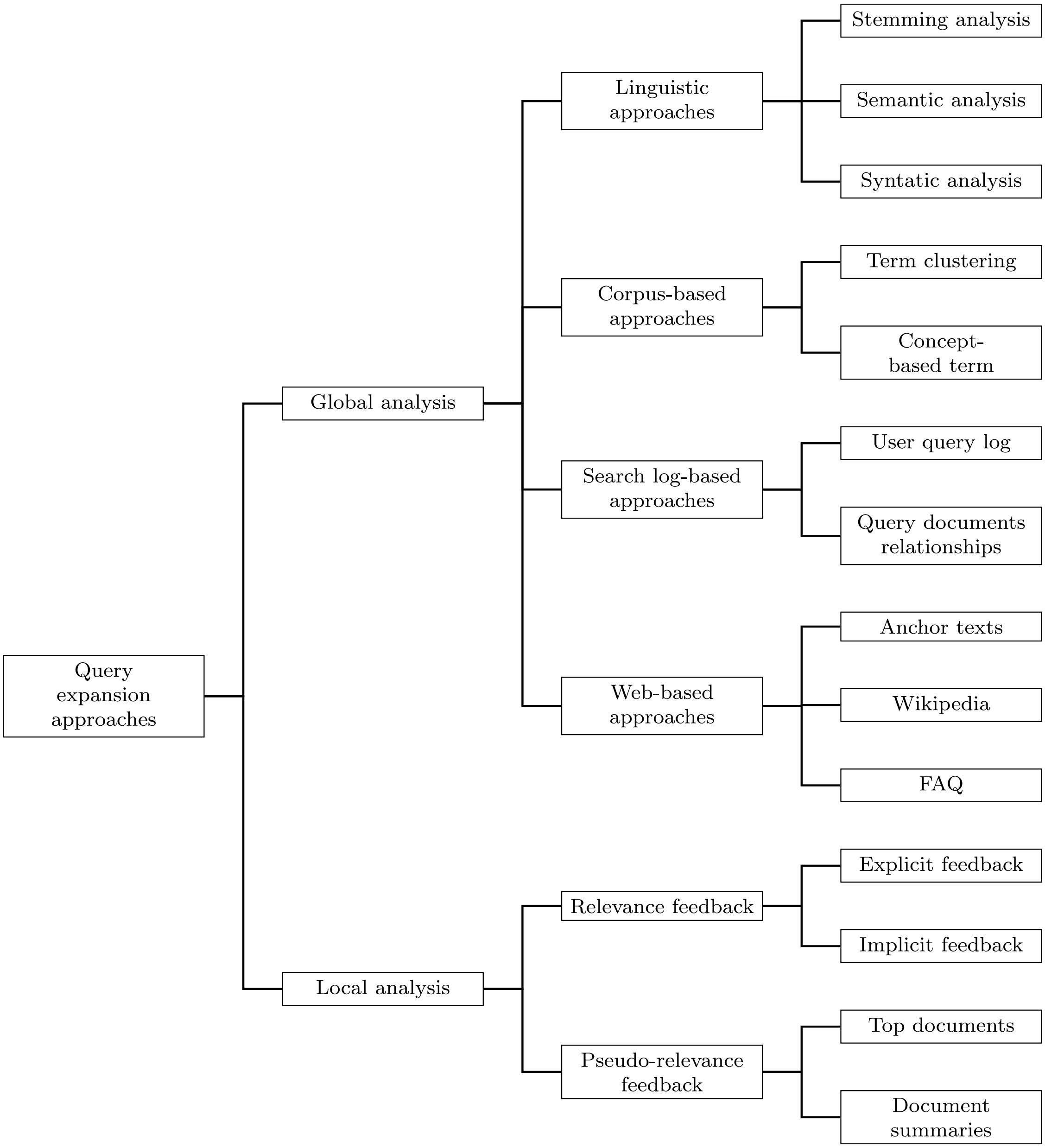

Fig. 5. Classification of QE approaches based on data sources.

sources: (i) linguistic-based, (ii) corpus-based, (iii) search log-based, and (iv) 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 thesauruses, 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 QE approaches in linguistic association to reduce the inflected word from its root word. The stemming algorithm (e.g., Porter, 1980) can be utilized either at retrieval time or at indexing time. When used during retrieval, terms from initially retrieved documents are picked, and then, these terms are harmonized with the morphological types of query terms (e.g., Krovetz, 1993; Paice, 1994). When used during indexing time, document word stems are picked, and then, these words are harmonized with the query root word stems (e.g., Hull et al., 1996). A morphological approach is an ordered way of studying the internal structure of the word. It has been shown to give better results than the stemming approach (Bilotti, Katz, & Lin, 2004; Moreau, Claveau, & Sébillot, 2007), however, it requires querying to be done in a structured way.

Use of semantic and contextual analysis are other popular QE approaches in the linguistic association. It includes knowledge sources such as Ontologies, LOD cloud, dictionaries and thesaurus. In the context of ontological based QE, Bhogal et al. (2007) use domain-specific and domain-independent ontologies. Wu, Ilyas, and Weddell (2011) utilize the rich semantics of domain ontology and evaluate the trade-off between the improvement in retrieval effectiveness and the computational cost. Several research works have been done on QE using a thesaurus. WordNet is a well-known thesaurus for expanding the initial query using word synsets. As discussed earlier, many of the research works use WordNet for expanding the initial query. For example, Voorhees (1994) uses WordNet to find the synonyms. Smeaton et al. (1995) use WordNet and POS tagger for expanding the initial query. However, this approach faces some practical problems, such as the absence of accurate matching between query and senses, the absence of proper nouns, and, one query term mapping to many noun synsets and collections. Generally, the utilization of WordNet for QE is beneficial only if the query words are unambiguous in nature (Gonzalo, Verdejo, Chugur, & Cigarran, 1998; Voorhees, 1994); using word sense disambiguation (WSD) to remove ambiguity is not easy (Navigli, 2009; Pal & Saha, 2015). Several research works have attempted to address the WSD problem. For example, Navigli and Velardi (2005) suggest that instead of considering the replacement of the initial query term with its synonyms, hyponyms, and hyperonyms, it is better to extract similar concepts from the query domain from WordNet (such as the common nodes and glossy terms). Gong et al. (2006) use the semantically similar information from WordNet present in different groups; this may be combined to expand the initial query. Zhang et al. (2009), Song et al. (2007), and Liu et al. (2004) combine WordNet concepts – that are extracted by applications of heuristic rules to match similar query terms – with other term extraction techniques. Shekarpour et al. (2013) use linguistic and semantic features of the initial query over linked data for QE as discussed earlier in Section 3.2.2. Recently, Agirre, de Lacalle, and Soroa (2014) introduced a WSD algorithm based on random walks over large Lexical Knowledge Bases (LKB). Their experiments give better results than other graph-based approaches when executed on a graph built from WordNet and eXtended WordNet (Mihalcea & Moldovan, 2001). Nowadays, Word Embeddings techniques are being widely used for QE, e.g., by Roy et al. (2016), Diaz et al. (2016) and Kuzi et al. (2016) as discussed earlier.

Another important approach that improves the linguistic information of the initial query is syntactic analysis (Zhang & Clark, 2011). Syntactic based QE uses the enhanced relational features of the query terms for expanding the initial query. It expands the query mostly through statistical approaches (Wu et al., 2011). It recognizes the term dependency statistically (Riezler et al., 2007) by employing techniques such as term co-occurrence. Sun et al. (2006) use this approach for extracting contextual terms and relations from the external corpus. Here, it uses two dependency relation based query expansion techniques for passage retrieval: Density-based system (DBS) and Relation based system (RBS). DBS makes use of relation analysis to extract high-quality contextual terms. RBS extracts relation paths for QE in a density and relation-based passage retrieval framework. The syntactic analysis approach may be beneficial for natural language queries in search tasks, where linguistic analysis can break the task into a sequence of decisions (Zhang & Clark, 2011) or integrate the taxonomic information effectively (Liu et al., 2008).

4.1.2. Corpus-based approaches

Corpus-based approaches examine the contents of the whole text corpus to recognize the expansion features to be utilized for QE. They are one of the earliest statistical approaches for QE. They create co-relations between terms based on co-occurrence statistics in the corpus to form sentences, paragraphs, or neighboring words, which are used in the expanded query. Corpus-based approaches have two admissible strategies: (1) term clustering (Crouch & Yang, 1992; Jones, 1971; Minker et al., 1972), which groups document terms into clusters based on their co-occurrences, and, (2) concept based terms (Fonseca et al., 2005; Natsev et al., 2007; Qiu & Frei, 1993), where expansion terms are based on the concept of query rather than the original query terms. Kuzi et al. (2016) select the expansion terms after the analysis of the corpus using word embeddings, where each term in the corpus is characterized by an embedded vector. Zhang et al. (2016) use four corpora as data sources (including one industry and three academic corpora) and present a Two-stage Feature Selection framework (TFS) for QE known as Supervised Query Expansion (SQE) (already discussed in Section 2.1.1). Some of the other approaches established an association thesaurus based on the whole corpus by using, e.g., context vectors Gauch et al. (1999), term co-occurrence (Carpineto et al., 2001), mutual information (Hu et al., 2006) and interlinked Wikipedia articles (Milne & Witten, 2008).

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 through the analysis of search logs. With the fast growing size of the web and the increasing use of web search engines, the abundance of search logs and their ease of use have made them an important source for QE. It usually contains user queries corresponding to the URLs of Web pages. Cui et al. (2002) use 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. The authors improved upon this in Cui et al. (2003) by using search logs for QE; their experiments give better results when compared with QE based on pseudo-relevance feedback. One of the advantages of using search logs is that it implicitly incorporates relevance feedback. On the other hand, it has been shown by White, Ruthven, and Jose (2005) that implicit measurements are relatively good, however, their performance may not be the same for all types of users and search tasks.

There are commonly two types of QE approaches used on the basis of web search logs. The first type considers queries as documents and extracts features of these queries that are related to the user's initial query (Huang, Chien, & Oyang, 2003). Among the techniques based on the first approach, some use their combined retrieval results (Huang & Efthimiadis, 2009), while some do not (e.g., Huang et al., 2003; Yin et al., 2009). In the second type of approach, the features are extracted on relational behavior of queries. For example, Baeza-Yates and Tiberi (2007) represent queries in a graph based vector space model (query-click bipartite graph) and analyze the graph constructed using the query logs. Cui et al. (2003), Riezler et al. (2007), and Cao, Jiang et al. (2008) extract the expansion terms directly from the clicked results. Fitzpatrick and Dent (1997), and Wang and Zhai (2007) use the top results from the past query terms entered by the users. Under the second approach, queries are also extracted from related documents (Billerbeck et al., 2003; Wang & Zhai, 2008) or through user clicks (Hua et al., 2013; Xue et al., 2004; Yin et al., 2009). The second type of approach is more popular and has been shown to give better results.

4.1.4. *Web-based approaches*

These approaches include Wikipedia and anchor texts from websites for expanding the user's original query. These approaches have gained popularity in recent times. Anchor text was first used by McBryan (1994) 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, an anchor text can play a role similar to the title since the anchor text pointing to a page can serve as a concise summary of its contents. It has been shown that user search queries and anchor texts are very similar because an anchor text is a brief characterization of its target page. Kraft and Zien (2004) used anchor texts for QE; their experimental results suggest that anchor texts can be used to improve the traditional QE based on query logs. On similar lines, Dang and Croft (2010) suggested that anchor texts can be an effective substitute for query logs. It demonstrated the effectiveness of QE techniques using log-based stemming through experiments on standard TREC collection dataset.

Another popular approach is the use of Wikipedia articles, titles and hyper-links (in-links and out-linsk) (ALMasri et al., 2013; Arguello et al., 2008). 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 QE. Recently, quite a few research works have used it for QE (e.g., Aggarwal & Buitelaar, 2012; ALMasri et al., 2013; Arguello et al., 2008; Li et al., 2007; Xu et al., 2009). Al-Shboul and Myaeng (2014) attempt to enrich initial queries using semantic annotations in Wikipedia articles combined with phrase-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 by (Karan & Šnajder, 2015) use domain-specific FAQs data for manual QE. Some of the other works using FAQs are Agichtein et al. (2004); Soricut and Brill (2006) and Riezler et al. (2007).

4.2. *Local analysis*

Local analysis includes QE techniques that select expansion terms from the collection of documents retrieved in response to the user's initial (unmodified) query. The working belief is that the documents 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 the user's original query: (1) Relevance feedback and (2) Pseudo-relevance feedback. These are discussed next.

4.2.1. *Relevance feedback (RF)*

In this approach, the user's feedback about documents retrieved in response to the initial query is collected; the feedback is about whether or not the retrieved documents are relevant to the user's query. The query is reformulated based on the documents found relevant as per the user's feedback. Rocchio's method (Rocchio, 1971) 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, the user explicitly evaluates the relevance of retrieved documents (as done in Harman, 1992; Salton & Buckley, 1990), whereas, in implicit feedback, the user's activity on the set of documents retrieved in response to the initial query is used to infer the user's preferences indirectly (e.g., as done in Chirita et al., 2007; Gao et al., 2015; Zhou et al., 2012). Relevance feedback suffers from the lack of semantics in the corpus (Wu et al., 2011). This restrains its applications in several occasions, for example, when the query concept is as general as a disjunction of more specific concepts (see Chapter 9 in the book by Manning et al., 2008). Some of the research works based on relevance feedback are Buckley, Salton, and Allan (1994); Ruthven and Lalmas (2003); Salton and Buckley (1997) and Manning et al. (2008); these have been discussed earlier in Section 2.2.3.

4.2.2. *Pseudo-relevance feedback (PRF)*

Here, neither explicit nor implicit feedback of the 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 QE. Pseudo-relevance feedback is also known as blind feedback, or, retrieval feedback. It has been discussed briefly earlier in Section 2.2.3. This technique was first proposed by Croft and Harper (1979), who employ this technique in a probabilistic model. Xu and Croft (2000) proposed “local context analysis” technique to extract the QE terms from the top documents retrieved in response to the initial query. Each of the candidate expansion terms is assigned a score on the basis of co-occurrence of query terms. The candidate terms with the highest score are selected for query reformulation. A recent work by Singh and Sharan (2016) uses fuzzy logic-based QE techniques and selects the top-retrieved documents based on PRF. Here, each expansion term is assigned a distinct relevance score using fuzzy rules. Then, the terms having the highest scores are selected for QE. The experimental results demonstrate that the proposed approach achieves significant improvement over individual expansion, expansion on the basis of the entire query and other related advanced methods. ALMasri, Berrut, and Chevallet (2016) proposed deep learning based QE technique and compared it with PRF and other expansion modules; the results show a notable improvement over other techniques using various language models for evaluation.

Considering the top retrieved documents 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 also be very similar. Hence, the expanded terms will not be useful for effective QE. Apart from using the top-ranked documents or their snippets, several other approaches have been proposed. For example, techniques based on passage extraction (Xu & Croft, 1996), text summarization (Lam-Adesina & Jones, 2001), and document summaries (Chang et al., 2006). Some of the other works using PRF are (Cao, Nie et al., 2008; Lv & Zhai, 2010; Xu et al., 2009); these have been discussed in earlier sections.

Comparative analysis: Of all the approaches mentioned earlier, corpus-based approaches are considered more effective than those based on linguistic-based approaches, whether it is global or local analysis. The main reason behind this is that linguistic-based approaches require a concrete linguistic relation (based on sense, meaning, concept etc.) between a query term and a relevant term for the latter to be discovered, while corpus-based approaches can discover the same relevant term simply based on co-occurrences with the query term. Generally, utilization of linguistic-based approaches is beneficial only if the query terms are unambiguous in nature. While several research works have attempted to remove the ambiguity using word sense disambiguation (WSD), exact solutions are very difficult to achieve. For statistical approaches, the local analysis seems to perform better than corpus-based approaches because the extracted features are query specific, whereas techniques based on Web data (such as user query logs or anchor texts) have not yet been systematically evaluated or compared with others on standard test collection.

The use of thesaurus, dictionaries, WordNet or ConceptNet presents some good expansion terms in linguistic-based approaches, but it also causes topic-drift. These approaches can only be used when we know the query’s domain or for domain-specific searches because domain-specific resources reduce the topic-drift in such cases. For local analysis, relevance feedback has been demonstrated to be more robust in performance than pseudo-relevance feedback. The primary reason behind this is that pseudo-relevance feedback depends significantly on the execution of the user’s initial query; if the initial query is poorly formulated or ambiguous, then the expansion terms extracted from the retrieved documents may not be relevant. Billerbeck and Zobel (2004) reported that pseudo-relevance feedback improved only one-third of queries in their experimental collection. However, based on the recent trends 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. Considering different data sources, an analysis specific to the local context of a data source can improve retrieval performance by combining global analysis with local feedback. Further, it can be concluded to use the phrase-based expansion technique that uses hybrid data sources for automatic query expansion. In addition to effectiveness and efficiency, we would like to highlight additional pointers for choosing QE technique: (1) Linguistic, Web and Search log-based approaches make use of data that are not always available or suitable for the information retrieval task, (2) Corpus-based approaches are not ideal for dynamic document collection, (3) Query-based approaches are dependent on the quality of the first-pass retrieval documents.

Many studies (Beaulieu, 1997; Koenemann & Belkin, 1996; Ruthven & Lalmas, 2003) have been done to compare the effectiveness of automatic and interactive query expansions. Intuitively, since the user is the one who decides which document is relevant to his/her query, the user should be able to make a better decision than the system with respect to the terms to be added to the initial query. However, experimental results do not offer conclusive results regarding interactive query expansion being more effective than automatic query expansion. For example, Beaulieu (1997) shows that automatic query expansion is more effective than interactive query expansion in the operational setting. On the other hand, Koenemann and Belkin (1996) reported user satisfaction and improved system performance with interactive query expansion system. However, Ruthven and Lalmas (2003) did a simulation study and reported that interactive query expansion has potential to achieve higher performance than automatic query expansion, but it is difficult for the users to filter the expansion terms that represent relevant documents.

In summary, there is a wide range of QE approaches that present various characteristics and are mostly useful or applicable in specific circumstances. The best option depends on the evaluation of several factors, including the type of queries, availability and features of external data sources, kind of collection being searched, facilities offered by the underlying weighting and ranking system, and efficiency requirements.

Table 8 summarizes applicability of QE techniques with respect to the approach and the Data sources used. Applicability and the data sources used by each technique have been presented in a comparative manner; all the techniques have been categorized under two main approaches: Global analysis and Local analysis.

Table 8
Applicability of QE techniques categorized with respect to data sources.

Approaches	Sub-Approaches	Data Sources used	Applicability	Publications
Global analysis	Linguistic approaches	Thesaurus, dictionaries, ontologies, LOD cloud, WordNet, ConceptNet	Word stemming, semantic and contextual analysis, syntactic analysis	Porter (1980), Krovetz (1993), Voorhees (1994), Bilotti et al. (2004), Sun et al. (2006), Bhogal et al. (2007), Zhang et al. (2009), Wu et al. (2011), Agirre et al. (2014), Kuzi et al. (2016)
	Corpus-based approaches	Corpus based thesaurus, text corpus	Term clustering, finding co-relation between terms, mutual information extraction, concept based term extraction	Jones (1971), Minker et al. (1972), Qiu and Frei (1993), Carpineto et al. (2001), Fonseca et al. (2005), Hu et al. (2006), Natsev et al. (2007), Kuzi et al. (2016)
	Search log-based approaches	Search logs, query logs, user logs	Features extraction based on relational behavior of user's queries, Query-Documents relationship	Cui et al. (2002, 2003), White et al. (2005), Wang and Zhai (2007), Yin et al. (2009), Hua et al. (2013)
	Web-based approaches	Wikipedia, anchor texts, FAQs	Enrich initial queries using semantic annotations, Associating hyper-links with linked pages, mutual QE	McBryan (1994), Kraft and Zien (2004), Li et al. (2007), Xu et al. (2009), ALMasri et al. (2013), Al-Shboul and Myaeng (2014), Karan and Šnajder (2015)
Local analysis	Relevance feedback	Retrieved documents based upon user's decision	Enrich user's query based on user's feedback	Rocchio (1971), Salton and Buckley (1990), Harman (1992), Salton and Buckley (1997), Chirita et al. (2007), Manning et al. (2008), Zhou et al. (2012), Gao et al. (2015)
	Pseudo-relevance feedback	Retrieved documents based upon top ranked documents	Enrich user's query based on top ranked documents (instead of user's feedback) retrieved in response to the initial query	Croft and Croft and Harper (1979), Xu and Croft (1996), Xu and Croft (2000), Lam-Adesina and Jones (2001), Chang et al. (2006), Cao, Nie et al. (2008), Lv and Zhai (2010), ALMasri et al. (2016), Singh and Sharan (2016)

5. Discussion and conclusions

This article has presented a comprehensive survey highlighting the current progress, emerging research directions, potential new research areas and novel classification of state-of-the-art approaches in the field of QE. The analysis was carried out over four key aspects: (1) data source, which is the collection of documents used for expanding the user's initial query, (2) working methodology, which describes the process for expanding the query, (3) importance and application, which discusses the importance of QE in IR and the use of this technique in the recent trend beyond the key area of IR, and (4) Core approaches, which discuss several QE approaches based on different features of data sources. Furthermore, this article presents a classification of QE approaches into two categories according to the various characteristics of data sources, namely: global analysis, and local analysis. Global analysis was further split into four subcategories: linguistic approaches, corpus-based approaches, search log-based approaches, and Web-based approaches. Local analysis was also split into two subcategories: relevance feedback and pseudo-relevance feedback.

Moreover, the survey provides a discussion of QE in the area of IR as well as the recent trends beyond the IR. QE can be defined as a process in IR that consists of choosing and adding expansion terms to the user's initial query with the goal of minimizing query-document mismatch to improve the retrieval performance. Although there is no perfect solution for the vocabulary mismatch problem in IR systems, QE has the capability to overcome the primary limitations. This is because QE provides the supporting explanation of the information needed for efficient IR, 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 the vocabulary mismatch problem. Over the past few decades, a lot of research has been done in the area of QE based on data sources used, applications, and expansion techniques. This article classifies the various data sources into four categories: documents used in the retrieval process, hand-built knowledge resources, external text collections and resources, and hybrid data sources. Recently, hybrid data sources have been used widely for QE; they are a combination of two or more data sources, more than often, web data being one of them. In research involving web data, Wikipedia is a popular data source because it is freely available and is the largest encyclopedia on the web, where articles are regularly updated, and new articles are 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 a remarkable improvement in retrieval effectiveness: both with respect to precision and recall. These results are a proof of the advancement of research in QE techniques.

With the ever growing wealth of information available on the Internet, web searching has become an integral part of our lives. Every web user wants personalized information according to their interests and commitments, and hence, IR systems need to personalize search results based on the query and the user's interests. For getting these results, IR systems need a personalized QE approach (QE approach based on personal preference), which learns and uses the user profile to reflect his interests as well as his intent. Such an approach can enhance the retrieval performance. We believe personalization of web search results will play an important part in QE research in future. In personalization of web searches, there are two things that should be taken into account. First, how information is presented or structured on the web, and second, how users interact with different personalized systems. This should affect the way a user-interface is designed so that it allows the system to learn more about the user by collecting information about him. Such collected-information will play an important role in QE for improving the IR.

Beyond the key area of IR, there are other recent applications where QE techniques are widely used. For example, Personalized Social Documents, Question Answering, Cross-Language Information Retrieval, Information Filtering, Multimedia Information Retrieval, Plagiarism Detection, Enterprise Search, Code Search, Biomedical IR, Classification of E-commerce, and Text Classification.

Finally, it can be said that after decades of research in QE, it has matured greatly. While challenges exist for further research, a lot of real life applications are using state of the art techniques. This article will hopefully help a researcher to better understand QE and its use in the area of IR.

Tables 9–11 summarize 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.

5.1. Epistemological genesis of this article

This section describes how the articles reviewed in this paper were discovered. Survey papers by Bhogal et al. (2007), Fu et al. (2005), Manning et al. (2008), Carpineto and Romano (2012), He and Ounis (2007), Biancalana et al. (2013), and Paik et al. (2014) were our getting started references. We were aware of these references through our prior exposure to the topic of our survey. Another thing we did while starting our review was to search for the term "Query Expansion" on Google Scholar. In the search results, by looking at the keywords noted in the relevant papers, we identified further related keywords, such as "Query Formulation","Information Retrieval", "Query Enhancement" and "Internet Searches". We searched these keywords on Google scholar looking for prominent papers related to query expansion; paying more attention to the papers published in the last 15–20 years.

Another common approach we took is: whenever we found an article, say *X*, to be an influential reference in the field, we also went through the papers that have cited *X* and the papers that have been cited by *X*. We applied this technique to our "getting started"

Table 9
Summary of research in the area of query expansion.

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

Table 10
Summary of Research in the area of QE (Cont. from Table 9).

Reference	Data sources	Term extraction methodology	Term representation	Term selection methodology	Weighting Schema
Lee et al. (2008)	Corpus & top-ranked documents	Clustering of top-ranked documents	Individual terms	Relevance model	Query language model
Cao, Nie et al. (2008)	Corpus & top-ranked documents	All terms in top retrieved documents	Individual terms	Term classification	Query language model
Arguello et al. (2008)	Wikipedia	Anchor texts in top retrieved Wikipedia documents	Phrases	Document rank & link frequency	Sum of entry likelihoods
Xu et al. (2009)	Wikipedia	All terms in top retrieved articles	Individual terms	Relevance model	Query language model
Yin et al. (2009)	Query logs & Snippets	All terms in top retrieved snippets & random walks on query-URL graph	Individual terms	Relevance model & mixture model	Query language model
Dang and Croft (2010)	Anchor texts	Adjacent terms in anchor text	Individual terms & Phrase	Kullback–Leibler divergence (KLD)	Substitution probability
Lv and Zhai (2010)	Corpus & top-ranked documents	All terms in top retrieved feedback documents	Individual terms	Positional relevance model (PRM)	Probabilistic reweighting
Kim et al. (2011)	Corpus	All terms in top retrieved documents	Individual terms	Decision tree-based	Boolean query
Bhatia et al. (2011)	Corpus	All terms in the corpus	Individual terms & Phrases	Document-centric approach	Correlation based weights
Miao et al. (2012)	Corpus	All Proximity-based terms in the corpus	Individual terms	Proximity-based feedback model (PRoc)	KLD scores
Zhou et al. (2012)	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)	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)	Wikipedia	In-link & out-link articles in top retrieved articles	Individual terms	Semantic similarity	Semantic similarity score
Pal et al. (2013)	Corpus	All terms in top retrieved documents	Individual terms	Association & distribution based term selection	KLD score
Bouchoucha et al. (2013)	ConceptNet	Diverse expansion terms from retrieved documents	Individual terms	MMRE (Maximal Marginal Relevance -based Expansion)	MMRE score
Augenstein et al. (2013)	LOD cloud	Neighbors term in whole graph of LOD	Individual terms	Mapping Keywords	tf-idf

Table 11
Summary of research in the area of QE (Cont. from Table 10).

Reference	Data sources	Term extraction methodology	Term representation	Term selection methodology	Weighting Schema
Pal et al. (2014)	WordNet	Synonyms, holonyms & meronyms of the query terms	Individual terms	Pseudo relevant documents	Normalized weights
Paik et al. (2014)	Top retrieved set of documents	All terms in feedback documents	Individual terms	Incremental Blind Feedback (IBF)	Scoring function
Al-Shboul and Myaeng (2014)	Wikipedia	Titles & in/out links of Wikipedia pages	Individual terms & Phrases	Wikipedia page surrogates	Phrase likelihood model
Dalton et al. (2014)	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)	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)	Freebase	All terms in top retrieved documents	Individual terms	tf.idf based PRF & Freebase's entity categories	tf.idf & Jensen-Shannon divergence score
Gan and Hong (2015)	Wikipedia & corpus	Top-k correlated terms in the corpus	Individual terms	Markov network model	Probabilistic query reweighting
Roy et al. (2016)	Corpus	All nearby terms at word embedding framework	Individual terms	K-nearest neighbor approach (KNN)	Query language model
Diaz et al. (2016)	Corpus	All terms in locally-trained word embeddings	Individual terms	local embedding approach	Query language models & KLD score
Singh and Sharan (2016)	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)	Corpora	Top-k terms in retrieved documents	Individual terms	Two-stage Feature Selection (TFS)	Probabilistic query reweighting
Zhou et al. (2017)	User profiles & folksonomy data	All semantically similar terms inside the enriched user profiles	Individual terms	Word embeddings & topic models	Topical weighting scheme

references, as well as to other prominent papers that we came across during the survey.

We also looked up digital libraries of the journals prominent in the field of "query expansion"/"information retrieval" such as Knowledge and Information Systems (KAIS), Information Retrieval Journal, ACM Computing Surveys, Information Processing and Management, and ACM Transactions on Information Systems. We did the same for prominent conferences such as Text REtrieval Conference (TREC), Knowledge Discovery and Data Mining (KDD), Conference and Labs of the Evaluation Forum (CLEF), Special Interest Group on Information Retrieval (SIGIR) and Forum for Information Retrieval Evaluation (FIRE).

Acknowledgments

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

References

Abdulla, A. A. A., Lin, H., Xu, B., & Banbhrani, S. K. (2016). Improving biomedical information retrieval by linear combinations of different query expansion techniques. *BMC Bioinformatics, 17*(7), 238.

Adriani, M., & Van Rijsbergen, C. (1999). *Term similarity-based query expansion for cross-language information retrieval. International conference on theory and practice of digital libraries*. Springer311–322.

Aggarwal, N., & Buitelaar, P. (2012). *Query expansion using Wikipedia and DBpedia. CLEF (online working notes/labs/workshop)*.

Agichtein, E., Lawrence, S., & Gravano, L. (2004). Learning to find answers to questions on the web. *ACM Transactions on Internet Technology (TOIT), 4*(2), 129–162.

Agirre, E., de Lacalle, O. L., & Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics, 40*(1), 57–84.

Agrawal, R., Imieliński, T., & Swami, A. (1993). *Mining association rules between sets of items in large databases. ACM SIGMOD record22. ACM SIGMOD record* ACM207–216.

Al-Shboul, B., & Myaeng, S.-H. (2014). Wikipedia-based query phrase expansion in patent class search. *Information Retrieval, 17*(5–6), 430–451.

Allan, J. (1996). *Incremental relevance feedback for information filtering. Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*. ACM270–278.

ALMasri, M., Berrut, C., & Chevallet, J.-P. (2013). *Wikipedia-based semantic query enrichment. Proceedings of the sixth international workshop on exploiting semantic annotations in information retrieval*. ACM5–8.

ALMasri, M., Berrut, C., & Chevallet, J.-P. (2016). *A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information. European conference on information retrieval*. Springer709–715.

Amati, G., Joost, C., & Rijsbergen, V. (2003). *Probabilistic models for information retrieval based on divergence from randomness*.

Amer, N. O., Mulhem, P., & Géry, M. (2016). *Toward word embedding for personalized information retrieval. NEU-IR: The SIGIR 2016 workshop on neural information retrieval*.

Anand, R., & Kotov, A. (2015). *An empirical comparison of statistical term association graphs with DBpedia and ConceptNet for query expansion. Proceedings of the 7th forum for information retrieval evaluation*. ACM27–30.

Arguello, J., Elsas, J. L., Callan, J., & Carbonell, J. G. (2008). Document representation and query expansion models for blog recommendation. *International Conference on Weblogs and Social Media, 2008*(0), 1.

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence, 31*(1), 132–164.

Attar, R., & Fraenkel, A. S. (1977). Local feedback in full-text retrieval systems. *Journal of the ACM (JACM), 24*(3), 397–417.

Augenstein, I., Gentile, A. L., Norton, B., Zhang, Z., & Ciravegna, F. (2013). *Mapping keywords to linked data resources for automatic query expansion. Extended semantic web conference*. Springer101–112.

Baeza-Yates, R., Hurtado, C., & Mendoza, M. (2004). *Query recommendation using query logs in search engines. International conference on extending database technology*. Springer588–596.

Baeza-Yates, R., & Tiberi, A. (2007). *Extracting semantic relations from query logs. Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM76–85.

Bai, J., Nie, J.-Y., & Cao, G. (2006). *Context-dependent term relations for information retrieval. Proceedings of the 2006 conference on empirical methods in natural language processing*. Association for Computational Linguistics551–559.

Bai, J., Nie, J.-Y., Cao, G., & Bouchard, H. (2007). *Using query contexts in information retrieval. Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval*. ACM15–22.

Bai, J., Song, D., Bruza, P., Nie, J.-Y., & Cao, G. (2005). *Query expansion using term relationships in language models for information retrieval. Proceedings of the 14th ACM international conference on information and knowledge management*. ACM688–695.

Ballesteros, L., & Croft, B. (1996). *Dictionary methods for cross-lingual information retrieval. International conference on database and expert systems applications*. Springer791–801.

Ballesteros, L., & Croft, W. B. (1997). *Phrasal translation and query expansion techniques for cross-language information retrieval. ACM SIGIR forum31. ACM SIGIR forum* ACM84–91.

Ballesteros, L., & Croft, W. B. (1998). *Resolving ambiguity for cross-language retrieval. Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*. ACM64–71.

Ballesteros, L. A. (2002). *Cross-language retrieval via transitive translation. Advances in information retrieval*. Springer203–234.

Barrington, L., Chan, A., Turnbull, D., & Lanckriet, G. (2007). *Audio information retrieval using semantic similarity. 2007 IEEE international conference on acoustics, speech and signal processing-ICASSP'072. 2007 IEEE international conference on acoustics, speech and signal processing-ICASSP'07* IEEEII–725.

Beaulieu, M. (1997). Experiments on interfaces to support query expansion. *Journal of Documentation, 53*(1), 8–19.

Beeferman, D., & Berger, A. (2000). *Agglomerative clustering of a search engine query log. Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining*. ACM407–416.

Belkin, N. J., & Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM, 35*(12), 29–38.

Bender, M., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J. X., ... Weikum, G. (2008). *Exploiting social relations for query expansion and result ranking. Data engineering workshop, 2008. ICDEW 2008. IEEE 24th international conference on*. IEEE501–506.

Bendersky, M., Metzler, D., & Croft, W. B. (2011). *Parameterized concept weighting in verbose queries. Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval*. ACM605–614.

Bernardini, A., & Carpineto, C. (2008). *Fub at TREC 2008 relevance feedback track: Extending rocchio with distributional term analysisTechnical Report*. DTIC Document.

Bhatia, S., Majumdar, D., & Mitra, P. (2011). *Query suggestions in the absence of query logs. Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval*. ACM795–804.

Bhattacharya, P., Goyal, P., & Sarkar, S. (2016). Using word embeddings for query translation for hindi to english cross language information retrieval. arXiv preprintarXiv:1608.01561.

Bhogal, J., Macfarlane, A., & Smith, P. (2007). A review of ontology based query expansion. *Information Processing & Management, 43*(4), 866–886.

Bian, J., Liu, Y., Agichtein, E., & Zha, H. (2008). *Finding the right facts in the crowd: Factoid question answering over social media. Proceedings of the 17th international conference on world wide web*. ACM467–476.

Biancalana, C., Gasparetti, F., Micarelli, A., & Sansonetti, G. (2013). Social semantic query expansion. *ACM Transactions on Intelligent Systems and Technology (TIST), 4*(4), 60.

Biancalana, C., & Micarelli, A. (2009). *Social tagging in query expansion: A new way for personalized web search. Computational science and engineering, 2009. CSE'09. international conference on4. Computational science and engineering, 2009. CSE'09. international conference on* IEEE1060–1065.

Billerbeck, B., Scholer, F., Williams, H. E., & Zobel, J. (2003). *Query expansion using associated queries. Proceedings of the twelfth international conference on information and knowledge management.* ACM2–9.

Billerbeck, B., & Zobel, J. (2004). *Questioning query expansion: An examination of behaviour and parameters. Proceedings of the 15th Australasian database conference-volume 27*. Australian Computer Society, Inc.69–76.

Bilotti, M. W., Katz, B., & Lin, J. (2004). *What works better for question answering: Stemming or morphological query expansion. Proceedings of the information retrieval for question answering (IR4QA) workshop at SIGIR2004. Proceedings of the information retrieval for question answering (IR4QA) workshop at SIGIR* 1–3.

de Boer, M., Schutte, K., & Kraaij, W. (2015). Knowledge based query expansion in complex multimedia event detection. *Multimedia Tools and Applications,* 1–19.

de Boer, M., Schutte, K., & Kraaij, W. (2016). Knowledge based query expansion in complex multimedia event detection. *Multimedia Tools and Applications, 75*(15), 9025–9043.

Borth, D., Ji, R., Chen, T., Breuel, T., & Chang, S.-F. (2013). *Large-scale visual sentiment ontology and detectors using adjective noun pairs. Proceedings of the 21st ACM international conference on multimedia*. ACM223–232.

Bouadjenek, M. R., Hacid, H., & Bouzeghoub, M. (Hacid, Bouzeghoub, 2013a). *Laicos: An open source platform for personalized social web search. Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining.* ACM1446–1449.

Bouadjenek, M. R., Hacid, H., & Bouzeghoub, M. (Hacid, Bouzeghoub, 2013b). *SOPRA: A new social personalized ranking function for improving web search. Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval.* ACM861–864.

Bouadjenek, M. R., Hacid, H., Bouzeghoub, M., & Daigremont, J. (2011). *Personalized social query expansion using social bookmarking systems. Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval.* ACM1113–1114.

Bouadjenek, M. R., Hacid, H., Bouzeghoub, M., & Vakali, A. (2016). Persador: Personalized social document representation for improving web search. *Information Sciences, 369*, 614–633.

Bouchoucha, A., He, J., & Nie, J.-Y. (2013). *Diversified query expansion using ConceptNet. Proceedings of the 22nd ACM international conference on conference on information & knowledge management.* ACM1861–1864.

Broder, A. (2002). *A taxonomy of web search. ACM SIGIR forum36. ACM SIGIR forum* ACM3–10.

Buckley, C., & Harman, D. (2004). *Reliable information access final workshop reportARDA Northeast Regional Research Center Technical Report 3*.

Buckley, C., Salton, G., & Allan, J. (1994). *The effect of adding relevance information in a relevance feedback environment. Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval.* Springer-Verlag New York, Inc.292–300.

Buckley, C., Salton, G., Allan, J., & Singhal, A. (1995). *Automatic query expansion using SMART: TREC 3. NIST special publication sp*69–80.

Büttcher, S., Clarke, C. L., & Cormack, G. V. (2016). *Information retrieval: Implementing and evaluating search engines.* MIT Press.

Cao, G., Gao, J., Nie, J.-Y., & Bai, J. (2007). *Extending query translation to cross-language query expansion with Markov chain models. Proceedings of the sixteenth ACM conference on conference on information and knowledge management*. ACM351–360.

Cao, G., Nie, J.-Y., Gao, J., & Robertson, S. (2008). *Selecting good expansion terms for pseudo-relevance feedback. Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval.* ACM243–250.

Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., et al. (2008). *Context-aware query suggestion by mining click-through and session data. Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining.* ACM875–883.

Carpineto, C., De Mori, R., Romano, G., & Bigi, B. (2001). An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems (TOIS), 19*(1), 1–27.

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR), 44*(1), 1.

Carpineto, C., Romano, G., & Giannini, V. (2002). Improving retrieval feedback with multiple term-ranking function combination. *ACM Transactions on Information Systems (TOIS), 20*(3), 259–290.

Cavalin, P., Figueiredo, F., de Bayser, M., Moyano, L., Candello, H., Appel, A., et al. (2016). *Building a question-answering corpus using social media and news articles. International conference on computational processing of the Portuguese language*. Springer353–358.

Chang, Y., Ounis, I., & Kim, M. (2006). Query reformulation using automatically generated query concepts from a document space. *Information Processing & Management, 42*(2), 453–468.

Chirita, P.-A., Firan, C. S., & Nejdl, W. (2007). *Personalized query expansion for the web. Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval.* ACM7–14.

Chu-Carroll, J., Prager, J., Czuba, K., Ferrucci, D., & Duboue, P. (2006). *Semantic search via XML fragments: A high-precision approach to ir. Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval.* ACM445–452.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22–29.

Cilibrasi, R. L., & Vitanyi, P. M. (2007). The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering, 19*(3).

Collins-Thompson, K. (2009). *Reducing the risk of query expansion via robust constrained optimization. Proceedings of the 18th ACM conference on information and knowledge management.* ACM837–846.

Collins-Thompson, K., & Callan, J. (2005). *Query expansion using random walk models. Proceedings of the 14th ACM international conference on information and knowledge management.* ACM704–711.

Collins-Thompson, K., & Callan, J. (2007). *Estimation and use of uncertainty in pseudo-relevance feedback. Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval.* ACM303–310.

Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F., & Voorhees, E. M. (2015). *TREC 2014 web track overviewTechnical Report.* DTIC Document.

Croft, B., & Lafferty, J. (2013). *Language modeling for information retrieval. 13.* Springer Science & Business Media.

Croft, W. B., & Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation, 35*(4), 285–295.

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). *Predicting query performance. Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval.* ACM299–306.

Crouch, C. J., & Yang, B. (1992). *Experiments in automatic statistical thesaurus construction. Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval.* ACM77–88.

Cui, H., Wen, J.-R., Nie, J.-Y., & Ma, W.-Y. (2002). *Probabilistic query expansion using query logs. Proceedings of the 11th international conference on world wide web.* ACM325–332.

Cui, H., Wen, J.-R., Nie, J.-Y., & Ma, W.-Y. (2003). Query expansion by mining user logs. *IEEE Transactions on Knowledge and Data Engineering, 15*(4), 829–839.

Dalton, J., & Dietz, L. (2013). *A neighborhood relevance model for entity linking. Proceedings of the 10th conference on open research areas in information retrieval.* Le Centre de Hautes Etudes Internationales D'informatique Documentaire149–156.

Dalton, J., Dietz, L., & Allan, J. (2014). *Entity query feature expansion using knowledge base links. Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval.* ACM365–374.

Dang, E. K. F., Luk, R. W., & Allan, J. (2016). A context-dependent relevance model. *Journal of the Association for Information Science and Technology, 67*(3), 582–593.

Dang, V., & Croft, B. W. (2010). *Query reformulation using anchor text. Proceedings of the third ACM international conference on web search and data mining.* ACM41–50.

Darwish, K., Magdy, W., et al. (2014). Arabic information retrieval. *Foundations and Trends® in Information Retrieval, 7*(4), 239–342.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR), 40*(2), 5.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodology),* 1–38.

Diaz, F. (2015). *Condensed list relevance models. Proceedings of the 2015 international conference on the theory of information retrieval.* ACM313–316.

Diaz, F., Mitra, B., & Craswell, N. (2016). Query expansion with locally-trained word embeddings. arXiv preprintarXiv:1605.07891.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*(3), 297–302.

Doszkocs, T. E. (1978). Aid, an associative interactive dictionary for online searching. *Online Review, 2*(2), 163–173.

Douze, M., Revaud, J., Schmid, C., & Jégou, H. (2013). *Stable hyper-pooling and query expansion for event detection. Proceedings of the IEEE international conference on computer vision*1825–1832.

Efthimiadis, E. N. (1996). Query expansion. *Annual Review of Information Science and Technology, 31*, 121–187.

Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS), 29*(2), 8.

Eguchi, K. (2005). *NTCIR-5 query expansion experiments using term dependence models. NTCIR*.

Eichstaedt, M., Patel, A. P., Lu, Q., Manber, U., & Rudkin, K. (2002). System and method for personalized information filtering and alert generation. US Patent 6,381, 594.

Fitzpatrick, L., & Dent, M. (1997). *Automatic feedback using past queries: Social searching? ACM SIGIR forum31. ACM SIGIR forum* ACM306–313.

Fonseca, B. M., Golgher, P., Pôssas, B., Ribeiro-Neto, B., & Ziviani, N. (2005). *Concept-based interactive query expansion. Proceedings of the 14th ACM international conference on information and knowledge management.* ACM696–703.

Franzoni, V. (2017). Just an update on pming distance for web-based semantic similarity in artificial intelligence and data mining. arXiv prepcrintarXiv:1701.02163.

Franzoni, V., & Milani, A. (2012). *PMING distance: A collaborative semantic proximity measure. Proceedings of the 2012 IEEE/WIC/ACM international joint conferences on web intelligence and intelligent agent technology-volume 02.* IEEE Computer Society442–449.

Fu, G., Jones, C. B., & Abdelmoty, A. I. (2005). *Ontology-based spatial query expansion in information retrieval. OTM confederated international conferences on the move to meaningful internet systems.* Springer1466–1482.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30*(11), 964–971.

Gaillard, B., Bouraoui, J. L., De Neef, E. G., & Boualem, M. (2010). *Query expansion for cross language information retrieval improvement. RCIS*337–342.

Gan, L., & Hong, H. (2015). Improving query expansion for information retrieval using Wikipedia. *International Journal of Database Theory and Application, 8*(3), 27–40.

Gao, Y., Xu, Y., & Li, Y. (2015). Pattern-based topics for document modelling in information filtering. *IEEE Transactions on Knowledge and Data Engineering, 27*(6), 1629–1642.

Gauch, S., Wang, J., & Rachakonda, S. M. (1999). A corpus analysis approach for automatic query expansion and its extension to multiple databases. *ACM Transactions on Information Systems (TOIS), 17*(3), 250–269.

Ghorab, M. R., Zhou, D., O'Connor, A., & Wade, V. (2013). Personalised information retrieval: Survey and classification. *User Modeling and User-Adapted Interaction, 23*(4), 381–443.

Gong, Z., Cheang, C. W., et al. (2006). *Multi-term web query expansion using WordNet. International conference on database and expert systems applications.* Springer379–388.

Gonzalo, J., Verdejo, F., Chugur, I., & Cigarran, J. (1998). Indexing with WordNet synsets can improve text retrieval. arXiv preprintcmp-lg/9808002.

Graupmann, J., Cai, J., & Schenkel, R. (2005). *Automatic query refinement using mined semantic relations. Web information retrieval and integration, 2005. WIRI'05. proceedings. international workshop on challenges in.* IEEE205–213.

Guisado-Gámez, J., Prat-Pérez, A., & Larriba-Pey, J. L. (2016). Query expansion via structural motifs in Wikipedia graph. arXiv preprintarXiv:1602.07217.

Hahm, G. J., Yi, M. Y., Lee, J. H., & Suh, H. W. (2014). A personalized query expansion approach for engineering document retrieval. *Advanced Engineering Informatics, 28*(4), 344–359.

Hanani, U., Shapira, B., & Shoval, P. (2001). Information filtering: Overview of issues, research and systems. *User modeling and user-adapted interaction, 11*(3), 203–259.

Harman, D. (1992). *Relevance feedback and other query modification techniques*.

Harman, D., & Voorhees, E. (1996). *Overview of the seventh text retrieval conference (TREC-7). Proceedings of the seventh text retrieval conference (TREC-7), NIST special publication 500-242*.

He, B., & Ounis, I. (2007). Combining fields for query expansion and adaptive query expansion. *Information processing & management, 43*(5), 1294–1307.

He, B., & Ounis, I. (2009). *Studying query expansion effectiveness. ECIR9. ECIR* Springer611–619.

Hersh, W., Price, S., & Donohoe, L. (2000). *Assessing thesaurus-based query expansion using the UMLS MNetathesaurus. Proceedings of the AMIA symposium.* American Medical Informatics Association344.

Hsu, M.-H., Tsai, M.-F., & Chen, H.-H. (2006). *Query expansion with ConceptNet and WordNet: An intrinsic comparison. Asia information retrieval symposium.* Springer1–13.

Hsu, M.-H., Tsai, M.-F., & Chen, H.-H. (2008). *Combining WordNet and ConceptNet for automatic query expansion: A learning approach. Asia information retrieval symposium.* Springer213–224.

Hu, J., Deng, W., & Guo, J. (2006). *Improving retrieval performance by global analysis. Pattern recognition, 2006. ICPR 2006. 18th international conference on2. Pattern recognition, 2006. ICPR 2006. 18th international conference on* IEEE703–706.

Hua, X.-S., Yang, L., Wang, J., Wang, J., Ye, M., Wang, K., ... Li, J. (2013). *Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines. Proceedings of the 21st ACM international conference on multimedia.* ACM243–252.

Huang, C.-K., Chien, L.-F., & Oyang, Y.-J. (2003). 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.

Huang, J., & Efthimiadis, E. N. (2009). *Analyzing and evaluating query reformulation strategies in web search logs. Proceedings of the 18th ACM conference on information and knowledge management.* ACM77–86.

Huber, S., Seiger, R., Kühnert, A., Theodorou, V., & Schlegel, T. (2016). Goal-based semantic queries for dynamic processes in the internet of things. *International Journal of Semantic Computing, 10*(02), 269–293.

Huber, S., Seiger, R., Kühnert, A., & Schlegel, T. (2016). *Using semantic queries to enable dynamic service invocation for processes in the internet of things. 2016 IEEE tenth international conference on semantic computing (ICSC).* IEEE214–221.

Hull, D. A., et al. (1996). Stemming algorithms: A case study for detailed evaluation. *Japan Analytical & Scientific Instruments Show, 47*(1), 70–84.

Imran, H., & Sharan, A. (2010). Selecting effective expansion terms for better information retrieval. *International Journal of Computer Science and Applications, Technomathematics Research Foundation, 7*(2), 52–64.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist, 11*(2), 37–50.

Jamil, H. M., & Jagadish, H. V. (2015). *A structured query model for the deep relational web. Proceedings of the 24th ACM international on conference on information and knowledge management.* ACM1679–1682.

Jammalamadaka, R. C., Salaka, V. K., Johnson, B. S., & King, T. H. (2015). Query expansion classifier for e-commerce. US Patent 9,135,330.

Jardine, N., & van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval, 7*(5), 217–240.

Jelinek, F. (1980). *Interpolated estimation of Markov source parameters from sparse data. Proc. workshop on pattern recognition in practice, 1980*.

Jelinek, F., & Mercer, R. L. (1980). *Interpolated estimation of Markov source parameters from sparse data. Proceedings of the workshop on pattern recognition in practice*.

Jian, F., Huang, J. X., Zhao, J., He, T., & Hu, P. (2016). *A simple enhancement for ad-hoc information retrieval via topic modelling. Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval.* ACM733–736.

Jones, K. S. (1971). *Automatic keyword classification for information retrieval*.

Jones, K. S., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and comparative experiments: Part 2. *Information Processing & Management, 36*(6), 809–840.

Jones, S., Gatford, M., Robertson, S., Hancock-Beaulieu, M., Secker, J., & Walker, S. (1995). Interactive thesaurus navigation: Intelligence rules OK? *Journal of the American Society for Information Science, 46*(1), 52.

Jourlin, P., Johnson, S. E., Jones, K. S., & Woodland, P. C. (1999). *General query expansion techniques for spoken document retrieval. ESCA tutorial and research workshop (ETRW) on accessing information in spoken audio*.

Junedi, M., Genevès, P., & Layaïda, N. (2012). *XML query-update independence analysis revisited. Proceedings of the 2012 ACM symposium on document engineering*. ACM95–98.

Kamps, J., Marx, M., Rijke, M.d., & Sigurbjörnsson, B. (2006). Articulating information needs in XML query languages. *ACM Transactions on Information Systems (TOIS), 24*(4), 407–436.

Kang, I.-H., & Kim, G. (2003). *Query type classification for web document retrieval. Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval*. ACM64–71.

Karan, M., & Šnajder, J. (2015). *Evaluation of manual query expansion rules on a domain specific FAQ collection. International conference of the cross-language evaluation forum for European languages*. Springer248–253.

Kato, M. P., Sakai, T., & Tanaka, K. (2012). *Structured query suggestion for specialization and parallel movement: Effect on search behaviors. Proceedings of the 21st international conference on world wide web*. ACM389–398.

Keyword (2018). *Query size by country* https://www.keyworddiscovery.com/keyword-stats.html

Khwileh, A., & Jones, G. J. (2016). *Investigating segment-based query expansion for user-generated spoken content retrieval. Content-based multimedia indexing (CBMI), 2016 14th international workshop on*. IEEE1–6.

Kim, Y., Seo, J., & Croft, W. B. (2011). *Automatic Boolean query suggestion for professional search. Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval*. ACM825–834.

Koenemann, J., & Belkin, N. J. (1996). *A case for interaction: A study of interactive information retrieval behavior and effectiveness. Proceedings of the SIGCHI conference on human factors in computing systems*. ACM205–212.

Kotov, A., & Zhai, C. (2012). *Tapping into knowledge base for concept feedback: Leveraging ConceptNet to improve search results for difficult queries. Proceedings of the fifth ACM international conference on web search and data mining*. ACM403–412.

Kraaij, W., Nie, J.-Y., & Simard, M. (2003). Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics, 29*(3), 381–419.

Kraft, R., & Zien, J. (2004). *Mining anchor text for query refinement. Proceedings of the 13th international conference on world wide web*. ACM666–674.

Krovetz, R. (1993). *Viewing morphology as an inference process. Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval*. ACM191–202.

Krovetz, R., & Croft, W. B. (1992). Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems (TOIS), 10*(2), 115–141.

Kumar, N., & Carterette, B. (2013). *Time based feedback and query expansion for twitter search. European conference on information retrieval*. Springer734–737.

Kuo, Y.-H., Chen, K.-T., Chiang, C.-H., & Hsu, W. H. (2009). *Query expansion for hash-based image object retrieval. Proceedings of the 17th ACM international conference on multimedia*. ACM65–74.

Kuzi, S., Shtok, A., & Kurland, O. (2016). *Query expansion using word embeddings. Proceedings of the 25th ACM international on conference on information and knowledge management*. ACM1929–1932.

Lam-Adesina, A. M., & Jones, G. J. (2001). *Applying summarization techniques for term selection in relevance feedback. Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*. ACM1–9.

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2007). *Light stemming for arabic information retrieval. Arabic computational morphology*. Springer221–243.

Latiri, C., Haddad, H., & Hamrouni, T. (2012). Towards an effective automatic query expansion process using an association rule mining approach. *Journal of Intelligent Information Systems, 39*(1), 209–247.

Lau, R. Y., Bruza, P. D., & Song, D. (2004). *Belief revision for adaptive information retrieval. Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval*. ACM130–137.

Lau, T., & Horvitz, E. (1999). *Patterns of search: Analyzing and modeling web query refinement. UM99 user modeling*. Springer119–128.

Lavrenko, V., & Allan, J. (2006). *Real-time query expansion in relevance modelsInternal Report no 473*. Center for Intelligent Information Retrieval-CIIR, University of Massachusetts.

Lavrenko, V., & Croft, W. B. (2001). *Relevance based language models. Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*. ACM120–127.

Lee, K. S., Croft, W. B., & Allan, J. (2008). *A cluster-based resampling method for pseudo-relevance feedback. Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval*. ACM235–242.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... Auer, S., et al. (2015). DBpedia–A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web, 6*(2), 167–195.

Lemos, O. A., de Paula, A. C., Zanichelli, F. C., & Lopes, C. V. (2014). *Thesaurus-based automatic query expansion for interface-driven code search. Proceedings of the 11th working conference on mining software repositories*. ACM212–221.

Levow, G.-A., Oard, D. W., & Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing & Management, 41*(3), 523–547.

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2*(1), 1–19.

Li, X., Uricchio, T., Ballan, L., Bertini, M., Snoek, C. G., & Bimbo, A. D. (2016). Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Computing Surveys (CSUR), 49*(1), 14.

Li, Y., Luk, W. P. R., Ho, K. S. E., & Chung, F. L. K. (2007). *Improving weak ad-hoc queries using Wikipedia asexternal corpus. Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval*. ACM797–798.

Lin, D., & Pantel, P. (2001). Discovery of inference rules for question-answering. *Natural Language Engineering, 7*(04), 343–360.

Lin, J., & Murray, G. C. (2005). *Assessing the term independence assumption in blind relevance feedback. Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*. ACM635–636.

Liu, D., Yan, S., Ji, R.-R., Hua, X.-S., & Zhang, H.-J. (2013). Image retrieval with query-adaptive hashing. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 9*(1), 2.

Liu, D.-R., Chen, Y.-H., Shen, M., & Lu, P.-J. (2015). 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.

Liu, H., & Singh, P. (2004). Conceptnet a practical commonsense reasoning tool-kit. *BT Technology Journal, 22*(4), 211–226.

Liu, S., Liu, F., Yu, C., & Meng, W. (2004). *An effective approach to document retrieval via utilizing WordNet and recognizing phrases. Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval*. ACM266–272.

Liu, X., Chen, F., Fang, H., & Wang, M. (2014). Exploiting entity relationship for query expansion in enterprise search. *Information Retrieval, 17*(3), 265–294.

Liu, Y., Li, C., Zhang, P., & Xiong, Z. (2008). *A query expansion algorithm based on phrases semantic similarity. Information processing (ISIP), 2008 international symposiums on*. IEEE31–35.

Lv, Y., & Zhai, C. (2009). *Positional language models for information retrieval. Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval*. ACM299–306.

Lv, Y., & Zhai, C. (2010). *Positional relevance model for pseudo-relevance feedback. Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval*. ACM579–586.

Lv, Y., Zhai, C., & Chen, W. (2011). *A boosting approach to improving pseudo-relevance feedback. Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval*. ACM165–174.

Macfarlane, A., Robertson, S., & Mccann, J. (1997). Parallel computing in information retrieval – An updated review. *Journal of Documentation, 53*(3), 274–315.

Magdy, W., & Jones, G. J. (2011). *A study on query expansion methods for patent retrieval. Proceedings of the 4th workshop on patent information retrieval.* ACM19–24.

Mahdabi, P., & Crestani, F. (2014). The effect of citation analysis on query expansion for patent retrieval. *Information Retrieval, 17*(5–6), 412–429.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* New York, NY, USA: Cambridge University Press.

Maron, M. (1965). *Mechanized documentation: The logic behind a probabilistic. Statistical association methods for mechanized documentation: Symposium proceedings269. Statistical association methods for mechanized documentation: Symposium proceedings* US Government Printing Office9.

Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM), 7*(3), 216–244.

McBryan, O. A. (1994). *GENVL and WWWW: Tools for taming the web. Proceedings of the first international world wide web conference, Geneva341*.

McNamee, P., & Mayfield, J. (2002). *Comparing cross-language query expansion techniques by degrading translation resources. Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval.* ACM159–166.

Metzler, D., & Croft, W. B. (2007). *Latent concept expansion using Markov random fields. Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval.* ACM311–318.

Miao, J., Huang, J. X., & Ye, Z. (2012). *Proximity-based rocchio's model for pseudo relevance. Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval.* ACM535–544.

Mihalcea, R., & Moldovan, D. I. (2001). *eXtended WordNet: Progress report. in proceedings of NAACL workshop on WordNet and other lexical resources.* Citeseer.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprintarXiv:1301.3781.

Mikroyannidis, A. (2007). Toward a social semantic web. *Computer, 40*(11), 113–115.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography, 3*(4), 235–244.

Milne, D., & Witten, I. H. (2008). *Learning to link with Wikipedia. Proceedings of the 17th ACM conference on information and knowledge management.* ACM509–518.

Minker, J., Wilson, G. A., & Zimmerman, B. H. (1972). An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval, 8*(6), 329–348.

Moldovan, D. I., & Mihalcea, R. (2000). Using WordNet and lexical operators to improve internet searches. *IEEE Internet Computing, 4*(1), 34.

Molino, P., Aiello, L. M., & Lops, P. (2016). Social question answering: Textual, user, and network features for best answer prediction. *ACM Transactions on Information Systems (TOIS), 35*(1), 4.

Montague, M., & Aslam, J. A. (2001). *Relevance score normalization for metasearch. Proceedings of the tenth international conference on information and knowledge management.* ACM427–433.

Moreau, F., Claveau, V., & Sébillot, P. (2007). *Automatic morphological query expansion using analogy-based machine learning. European conference on information retrieval.* Springer222–233.

Mulhem, P., Amer, N. O., & Géry, M. (2016). Axiomatic term-based personalized query expansion using bookmarking system. In S. Hartmann, & H. Ma (Eds.). *Database and expert systems applications: 27th international conference, DEXA 2016, Porto, Portugal, September 5–8, 2016, Proceedings, Part II* (pp. 235–243)). Cham: Springer International Publishing.

Natsev, A. P., Haubold, A., Tešić, J., Xie, L., & Yan, R. (2007). *Semantic concept-based query expansion and re-ranking for multimedia retrieval. Proceedings of the 15th ACM international conference on multimedia.* ACM991–1000.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR), 41*(2), 10.

Navigli, R., & Velardi, P. (2005). Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(7), 1075–1086.

Nawab, R. M. A., Stevenson, M., & Clough, P. (2016). An IR-Based approach utilizing query expansion for plagiarism detection in MEDLINE. *IEEE/ACM transactions on computational biology and bioinformatics, 14*(4), 796–804.

Nie, J.-Y., Simard, M., Isabelle, P., & Durand, R. (1999). *Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval.* ACM74–81.

Nie, L., Jiang, H., Ren, Z., Sun, Z., & Li, X. (2016). Query expansion based on crowd knowledge for code search. *IEEE Transactions on Services Computing, 9*(5), 771–783.

Paice, C. D. (1994). *An evaluation method for stemming algorithms. Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval.* Springer-Verlag New York, Inc.42–50.

Paik, J. H., Pal, D., & Parui, S. K. (2014). Incremental blind feedback: An effective approach to automatic query expansion. *ACM Transactions on Asian Language Information Processing (TALIP), 13*(3), 13.

Pakhomov, S. V., Finley, G., McEwan, R., Wang, Y., & Melton, G. B. (2016). Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics, 32*(23), 3635–3644.

Pal, A. R., & Saha, D. (2015). Word sense disambiguation: A survey. arXiv preprintarXiv:1508.01346.

Pal, D., Mitra, M., & Bhattacharya, S. (2015). Exploring query categorisation for query expansion: A study. arXiv preprintarXiv:1509.05567.

Pal, D., Mitra, M., & Datta, K. (2013). Query expansion using term distribution and term association. arXiv preprintarXiv:1303.0667.

Pal, D., Mitra, M., & Datta, K. (2014). Improving query expansion using WordNet. *Journal of the Association for Information Science and Technology, 65*(12), 2469–2478.

Pane, J. F., & Myers, B. A. (2000). *Improving user performance on Boolean queries. Chi'00 extended abstracts on human factors in computing systems.* ACM269–270.

Panovich, K., Miller, R., & Karger, D. (2012). *Tie strength in question & answer on social network sites. Proceedings of the ACM 2012 conference on computer supported cooperative work.* ACM1057–1066.

Peat, H. J., & Willett, P. (1991). 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.

Pirkola, A., Hedlund, T., Keskustalo, H., & Järvelin, K. (2001). Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval, 4*(3–4), 209–230.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Porter, M. F. (1982). Implementing a probabilistic information retrieval system. *Information Technology: Research and Development, 1*(2), 131–156.

Pound, J., Ilyas, I. F., & Weddell, G. (2010). *Expressive and flexible access to web-extracted data: A keyword-based structured query language. Proceedings of the 2010 ACM SIGMOD international conference on management of data.* ACM423–434.

Qiu, Y., & Frei, H.-P. (1993). *Concept based query expansion. Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval.* ACM160–169.

Radwan, K. (1994). *Vers l'acces multilingue en langage naturel aux bases de donnees textuelles* Ph.D. thesis.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. arXiv preprintarxiv:cmp-lg/9511007.

Riezler, S., Liu, Y., & Vasserman, A. (2008). *Translating queries into snippets for improved query expansion. Proceedings of the 22nd international conference on computational linguistics-volume 1.* Association for Computational Linguistics737–744.

Riezler, S., Vasserman, A., Tsochantaridis, I., Mittal, V., & Liu, Y. (2007). *Statistical machine translation for query expansion in answer retrieval. Annual meeting-association for computational linguistics45. Annual meeting-association for computational linguistics* 464.

van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation, 33*(2), 106–119.

Rijsbergen, C. J. V. (1979). *Information retrieval* (2nd ed.). Newton, MA, USA: Butterworth-Heinemann.

Rivas, A. R., Iglesias, E. L., & Borrajo, L. (2014). Study of query expansion techniques and their application in the biomedical information retrieval. *The Scientific World Journal, 2014*.

Robertson, A. M., & Willett, P. (1993). A comparison of spelling-correction methods for the identification of word forms in historical text databases. *Literary and Linguistic Computing, 8*(3), 143–152.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation, 60*(5), 503–520.

Robertson, S. E. (1990). On term selection for query expansion. *Journal of Documentation, 46*(4), 359–364.

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science, 27*(3), 129–146.

Robertson, S. E., & Walker, S. (1994). *Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval.* Springer-Verlag New York, Inc.232–241.

Robertson, S. E., & Walker, S. (2000). *Microsoft cambridge at TREC-9: Filtering track. TREC*.

Robertson, S. E., Walker, S., Beaulieu, M., & Willett, P. (1999). *Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive track. NIST Special Publication SP (500)* 253–264.

Rocchio, J. J. (1971). *Relevance feedback in information retrieval*.

Roy, D., Paul, D., Mitra, M., & Garain, U. (2016). Using word embeddings for automatic query expansion. arXiv preprintarXiv:1606.07608.

Ruthven, I., & Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review, 18*(2), 95–145.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer.* Reading: Addison-Wesley.

Salton, G. (1991). Developments in automatic text retrieval. *Science, 253*(5023), 974–980.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science, 41*, 288–297.

Salton, G., & Buckley, C. (1997). Improving retrieval performance by relevance feedback. *Readings in Information Retrieval, 24*(5), 355–363.

Savoy, J. (2005). 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.

Shah, C., & Croft, W. B. (2004). *Evaluating high accuracy retrieval techniques. Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval.* ACM2–9.

Shekarpour, S., Höffner, K., Lehmann, J., & Auer, S. (2013). *Keyword query expansion on linked data using linguistic and semantic features. Semantic computing (ICSC), 2013 IEEE seventh international conference on.* IEEE191–197.

Sheridan, P., & Ballerini, J. P. (1996). *Experiments in multilingual information retrieval using the spider system. Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval.* ACM58–65.

Sihvonen, A., & Vakkari, P. (2004). Subject knowledge improves interactive query expansion assisted by a thesaurus. *Journal of Documentation, 60*(6), 673–690.

Singh, J., Prasad, M., Prasad, O. K., Joo, E. M., Saxena, A. K., & Lin, C.-T. (2016). A novel fuzzy logic model for pseudo-relevance feedback-based query expansion. *International Journal of Fuzzy Systems, 18*(6), 980–989.

Singh, J., & Sharan, A. (2016). A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach. *Neural Computing and Applications,* 1–24.

Singhal, A., & Pereira, F. (1999). *Document expansion for speech retrieval. Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval.* ACM34–41.

Smeaton, A. F., Kelledy, F., & O'Donnell, R. (1995). *TREC-4 experiments at Dublin City University: Thresholding posting lists, query expansion with WordNet and pos tagging of Spanish. TREC*373–389.

Song, M., Song, I.-Y., Hu, X., & Allen, R. B. (2007). Integration of association rules and ontologies for semantic query expansion. *Data & Knowledge Engineering, 63*(1), 63–75.

Song, R., Yu, L., Wen, J.-R., & Hon, H.-W. (2011). *A proximity probabilistic model for information retrievalTechnical Report.* Microsoft Research.

Soricut, R., & Brill, E. (2006). Automatic question answering using the web: Beyond the factoid. *Information Retrieval, 9*(2), 191–206.

Spink, A., Wolfram, D., Jansen, M. B., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology, 52*(3), 226–234.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). *Short text classification in twitter to improve information filtering. Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval.* ACM841–842.

Statista (2017). Average number of search terms for online search queries in the United States as of August 2017. https://www.statista.com/statistics/269740/number-of-search-terms-in-internet-research-in-the-us/.

Stokoe, C., Oakes, M. P., & Tait, J. (2003). *Word sense disambiguation in information retrieval revisited. Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval.* ACM159–166.

Sun, R., Ong, C.-H., & Chua, T.-S. (2006). *Mining dependency relations for query expansion in passage retrieval. Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval.* ACM382–389.

Tejedor, J., Fapšo, M., Szöke, I., Černockỳ, J., & Grézl, F. (2012). Comparison of methods for language-dependent and language-independent query-by-example spoken term detection. *ACM Transactions on Information Systems (TOIS), 30*(3), 18.

Tellex, S., Kollar, T., Shaw, G., Roy, N., & Roy, D. (2010). *Grounding spatial language for video search. International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction.* ACM31.

Thomas, S. S., Gupta, S., & Venkatesh, K. (2016). Perceptual synoptic view-based video retrieval using metadata. *Signal, Image and Video Processing,* 1–7.

Turtle, H. (1994). *Natural language vs. Boolean query evaluation: A comparison of retrieval performance. Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval.* Springer-Verlag New York, Inc212–220.

Unger, C., Ngomo, A.-C. N., & Cabrio, E. (2016). *6th open challenge on question answering over linked data (QALD-6). Semantic web evaluation challenge.* Springer171–177.

Vaidya, J., & Clifton, C. (2002). *Privacy preserving association rule mining in vertically partitioned data. Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining.* ACM639–644.

Van Rijsbergen, C. J. (1986). A non-classical logic for information retrieval. *The Computer Journal, 29*(6), 481–485.

Voorhees, E. M. (1994). *Query expansion using lexical-semantic relations. SIGIR94.* Springer61–69.

Wang, F., & Lin, L. (2016). *Domain lexicon-based query expansion for patent retrieval. Natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD), 2016 12th international conference on.* IEEE1543–1547.

Wang, P., Xu, B., Xu, J., Tian, G., Liu, C.-L., & Hao, H. (2016). Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing, 174*, 806–814.

Wang, X., & Zhai, C. (2007). *Learn from web search logs to organize search results. Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval.* ACM87–94.

Wang, X., & Zhai, C. (2008). *Mining term association patterns from search logs for effective query reformulation. Proceedings of the 17th ACM conference on information and knowledge management.* ACM479–488.

Wei, X., & Croft, W. B. (2007). *Modeling term associations for ad-hoc retrieval performance within language modeling framework. European conference on information retrieval.* Springer52–63.

Wen, J.-R., Nie, J.-Y., & Zhang, H.-J. (2002). Query clustering using user logs. *ACM Transactions on Information Systems, 20*(1), 59–81.

White, R. W., Ruthven, I., & Jose, J. M. (2005). *A study of factors affecting the utility of implicit relevance feedback. Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval.* ACM35–42.

Willett, P. (1988). Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management, 24*(5), 577–597.

Wong, W., Luk, R. W. P., Leong, H. V., Ho, K., & Lee, D. L. (2008). Re-examining the effects of adding relevance information in a relevance feedback environment. *Information Processing & Management, 44*(3), 1086–1116.

Wu, H., & Fang, H. (2013). *An incremental approach to efficient pseudo-relevance feedback. Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval.* ACM553–562.

Wu, H., Wu, W., Zhou, M., Chen, E., Duan, L., & Shum, H.-Y. (2014). *Improving search relevance for short queries in community question answering. Proceedings of the 7th ACM international conference on web search and data miningWSDM '14*New York, NY, USA: ACM43–52.

Wu, J., Ilyas, I., & Weddell, G. (2011). *A study of ontology-based query expansionTechnical report CS-2011–04*.

Wu, Y., Liu, X., Xie, M., Ester, M., & Yang, Q. (2016). *Cccf: Improving collaborative filtering via scalable user-item co-clustering. Proceedings of the ninth ACM international* *conference on web search and data mining*. ACM73–82.

Wu, Z., & Palmer, M. (1994). *Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on association for computational linguistics*. Association for Computational Linguistics133–138.

Xie, H., Zhang, Y., Tan, J., Guo, L., & Li, J. (2014). Contextual query expansion for image retrieval. *IEEE Transactions on Multimedia, 16*(4), 1104–1114.

Xiong, C., & Callan, J. (2015). *Query expansion with freebase. Proceedings of the 2015 international conference on the theory of information retrieval*. ACM111–120.

Xu, J., & Croft, W. B. (1996). *Query expansion using local and global document analysis. Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*. ACM4–11.

Xu, J., & Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems, 18*(1), 79–112.

Xu, Y., Jones, G. J., & Wang, B. (2009). *Query dependent pseudo-relevance feedback based on Wikipedia. Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval*. ACM59–66.

Xue, G.-R., Zeng, H.-J., Chen, Z., Yu, Y., Ma, W.-Y., Xi, W., & Fan, W. (2004). *Optimizing web search using web click-through data. Proceedings of the thirteenth ACM international conference on information and knowledge management*. ACM118–126.

Yao, Y., Yi, J., Liu, Y., Zhao, X., & Sun, C. (2015). *Query processing based on associated semantic context inference. Information science and control engineering (ICISCE), 2015 2nd international conference on*. IEEE395–399.

Yin, Z., Shokouhi, M., & Craswell, N. (2009). *Query expansion using external evidence. European conference on information retrieval*. Springer362–374.

Yu, C. T., Buckley, C., Lam, K., & Salton, G. (1983). *A generalized term dependence model in information retrievalTechnical Report*. Cornell University.

Yu, K., Tresp, V., & Yu, S. (2004). *A nonparametric hierarchical bayesian framework for information filtering. Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval*. ACM353–360.

Zervakis, L., Tryfonopoulos, C., Skiadopoulos, S., & Koubarakis, M. (2017). Query reorganisation algorithms for efficient Boolean information filtering. *IEEE Transactions on Knowledge & Data Engineering*(1) 1–1

Zhai, C., & Lafferty, J. (2001). *Model-based feedback in the language modeling approach to information retrieval. Proceedings of the tenth international conference on information and knowledge management*. ACM403–410.

Zhang, C.-J., & Zeng, A. (2012). Behavior patterns of online users and the effect on information filtering. *Physica A: Statistical Mechanics and its Applications, 391*(4), 1822–1830.

Zhang, J., Deng, B., & Li, X. (2009). *Concept based query expansion using WordNet. Proceedings of the 2009 international e-conference on advanced science and technology*. IEEE Computer Society52–55.

Zhang, Y., & Clark, S. (2011). Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics, 37*(1), 105–151.

Zhang, Z., Wang, Q., Si, L., & Gao, J. (2016). *Learning for efficient supervised query expansion via two-stage feature selection. Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval*. ACM265–274.

Zhong, Z., & Ng, H. T. (2012). *Word sense disambiguation improves information retrieval. Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1*. Association for Computational Linguistics273–282.

Zhou, D., Lawless, S., Liu, J., Zhang, S., & Xu, Y. (2015). *Query expansion for personalized cross-language information retrieval. Semantic and social media adaptation and personalization (SMAP), 2015 10th international workshop on*. IEEE1–5.

Zhou, D., Lawless, S., & Wade, V. (2012). Improving search via personalized query expansion using social media. *Information Retrieval, 15*(3–4), 218–242.

Zhou, D., Lawless, S., Wu, X., Zhao, W., Liu, J., & Lewandowski, D. (2016). A study of user profile representation for personalized cross-language information retrieval. *ASLIB Journal of Information Management, 68*(4).

Zhou, D., Wu, X., Zhao, W., Lawless, S., & Liu, J. (2017). Query expansion with enriched user profiles for personalized search utilizing folksonomy data. *IEEE Transactions on Knowledge and Data Engineering*.

Zimmer, C., Tryfonopoulos, C., & Weikum, G. (2008). *Exploiting correlated keywords to improve approximate information filtering. Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval*. ACM323–330.

Zingla, M. A., Chiraz, L., & Slimani, Y. (2016). Short query expansion for microblog retrieval. *Procedia Computer Science, 96*, 225–234.