A Survey of Automatic Query Expansion in Information Retrieval

CLAUDIO CARPINETO and GIOVANNI ROMANO, Fondazione Ugo Bordoni

The relative ineffectiveness of information retrieval systems is largely caused by the inaccuracy with which a query formed by a few keywords models the actual user information need. One well known method to overcome this limitation is automatic query expansion (AQE), whereby the user's original query is augmented by new features with a similar meaning. AQE has a long history in the information retrieval community but it is only in the last years that it has reached a level of scientific and experimental maturity, especially in laboratory settings such as TREC. This survey presents a unified view of a large number of recent approaches to AQE that leverage various data sources and employ very different principles and techniques. The following questions are addressed. Why is query expansion so important to improve search effectiveness? What are the main steps involved in the design and implementation of an AQE component? What approaches to AQE are available and how do they compare? Which issues must still be resolved before AQE becomes a standard component of large operational information retrieval systems (e.g., search engines)?

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms: Algorithms, Experimentation, Measurement, Performance

Additional Key Words and Phrases: Query expansion, query refinement, search, word associations, pseudo-relevance feedback, document ranking

ACM Reference Format:
Carpineto, C. and Romano, G. 2012. A survey of automatic query expansion in information retrieval. ACM Comput. Surv. 44, 1, Article 1 (January 2012), 50 pages.
DOI = 10.1145/2071389.2071390 http://doi.acm.org/10.1145/2071389.2071390

1. INTRODUCTION

Current information retrieval systems, including Web search engines, have a standard interface consisting of a single input box that accepts keywords. The keywords submitted by the user are matched against the collection index to find the documents that contain those keywords, which are then sorted by various methods. When a user query contains multiple topic-specific keywords that accurately describe his information need, the system is likely to return good matches; however, given that user queries are usually short and that the natural language is inherently ambiguous, this simple retrieval model is in general prone to errors and omissions.

The most critical language issue for retrieval effectiveness is the term mismatch problem: the indexers and the users do often not use the same words. This is known as the *vocabulary problem* Furnas et al. [1987], compounded by synonymy (same word with different meanings, such as "java") and polysemy (different words with the same

Authors' address: C. Carpineto and G. Romano, Fondazione Ugo Bordoni, Via Baldassarre Castiglione 59, 00142, Rome, Italy; email: carpinet@fub.it.

DOI 10.1145/2071389.2071390 http://doi.acm.org/10.1145/2071389.2071390

or similar meanings, such as "tv" and "television"). Synonymy, together with word inflections (such as with plural forms, "television" versus "televisions"), may result in a failure to retrieve relevant documents, with a decrease in *recall* (the ability of the system to retrieve all relevant documents). Polysemy may cause retrieval of erroneous or irrelevant documents, thus implying a decrease in *precision* (the ability of the system to retrieve only relevant documents).

To deal with the vocabulary problem, several approaches have been proposed including interactive query refinement, relevance feedback, word sense disambiguation, and search results clustering. One of the most natural and successful techniques is to expand the original query with other words that best capture the actual user intent, or that simply produce a more useful query—a query that is more likely to retrieve relevant documents. Automatic query expansion (AQE) has a long history in information retrieval (IR), as it has been suggested as early as 1960 by Maron and Kuhns [1960]. Early work investigated a range of seminal techniques that have been subsequently improved and extended in various ways, for example, vector feedback [Ide 1971; Rocchio 1971], term-term clustering [Harper and van Rijsbergen 1978; Lesk 1969; Minker et al. 1972], and comparative analysis of term distributions [Doszkocs 1978; Porter 1982]. On the other hand, in a number of early experiments performed on small scale collections inconclusive results were achieved about the retrieval effectiveness of such techniques, with gain in recall often compensated by the corresponding loss in precision (see Salton and Buckley [1990] and Harman [1992] for a review).

As the volume of data has dramatically increased while the number of searcher-supplied query terms has remained very low, research on AQE has been revamped. Web search is the best case in point. According to *Hitwise*[1], in 2009 the average query length was 2.30 words, the same as that reported ten years before in Lau and Horvitz [1999]. While there has been a slight increase in the number of long queries (of five or more words), the most prevalent queries are still those of one, two, and three words. In this situation, the vocabulary problem has become even more serious because the paucity of query terms reduces the possibility of handling synonymy while the heterogeneity and size of data make the effects of polysemy more severe. The need for and the scope of AQE have thus increased.

In the last years, a huge number of AQE techniques have been presented using a variety of approaches that leverage on several data sources and employ sophisticated methods for finding new features correlated with the query terms. Today, there are firmer theoretical foundations and a better understanding of the utility and limitations of AQE; e.g., which are the critical parameters affecting the method performance, what type of queries is AQE useful for, and so on. At the same time, the basic techniques are being increasingly used in conjunction with other mechanisms to increase their effectiveness, including method combination, more active selection of information sources, and discriminative policies of method application. These scientific advances have been corroborated by very positive experimental findings obtained in laboratory settings. In fact, AQE has regained much popularity thanks to the evaluation results obtained at the Text REtrieval Conference series (TREC)[2], where most participants have made use of this technique, reporting noticeable improvements in retrieval performance.

AQE is currently considered an extremely promising technique to improve the retrieval effectiveness of document ranking and there are signs that it is being adopted

[1]http://www.hitwise.com/us/press-center/press-releases/2009/google-searches-oct-09/
[2]http://trec.nist.gov/

in commercial applications, especially for desktop and intranet searches. For instance, Google Enterprise, MySQL, and Lucene provide the user with an AQE facility that can be turned on or off. In contrast, it has not yet been regularly employed in the major operational Web IR systems such as search engines.

There are several explanations for the limited uptake of AQE in Web search. First, the fast response times required by Web search applications may prevent the use of some computationally expensive AQE techniques. Second, current AQE techniques are optimized to perform well on average, but are unstable and may cause degradation of search service for some queries. Also, the emphasis of AQE on improving recall (as opposed to guaranteeing high precision) is less important, given that there is usually an abundance of relevant documents and that many users look only at the first page of results. Third, there is probably an issue with the acceptance of AQE, due to the limited usability and transparency of an IR system implementing AQE: the user may get confused if the system retrieves documents that do not contain the original query terms. On the other hand, these features are less important in many other IR applications (e.g., search by experts in specialized domains), where a straightforward application of AQE may have no major contraindications. One of the objectives of this survey is to critically assess the performance limitations of this technique and discuss what we need to push it forward.

Although AQE has received a great deal of attention in the recent literature on IR and search, very little work has been done to review such studies. One notable exception is Bhogal et al. [2007], which however reviews a specific approach to AQE: using ontologies. AQE has also been covered in the books Baeza-Yates and Ribeiro-Neto [1999] and Manning et al. [2008], with a focus on early techniques for finding term correlations, and has a dedicated entry in the Encyclopedia of Database Systems Vechtomova [2009]. This article is the first comprehensive study of AQE that deals with all processing steps, reviews the major techniques including the recent ones, discusses their retrieval performance, identifies open issues, and suggests research directions.

After discussing how AQE can improve not only recall but also precision, we describe the main computational steps involved, from data acquisition and preprocessing, to candidate feature generation and ranking, to feature selection, and finally to query reformulation. This modelization accounts for a large number of proposed approaches, with each approach usually fitting in one or more sections of the full processing pipeline. Besides summarizing current practice, it can be used as a blueprint for designing and implementing an AQE component for a ranking system. We also provide a classification of existing techniques that is more oriented towards methodological aspects; e.g., the source of data, the feature extraction method, and the representation of the expanded query. The latter characterization is more useful for system comparison.

The remainder of the article has the following organization. We first provide a pragmatic definition of AQE (Section 2), discuss why and under which assumptions it produces more accurate results than using unexpanded queries (Section 3), and briefly review other applications of AQE in addition to document ranking (Section 4) and different approaches to the vocabulary problem (Section 5). Then, in Section 6, we describe how AQE works, identifying the main computational steps in which the whole process can be broken down. Section 7 is devoted to a classification of existing approaches: we provide a broad taxonomy by data source and by expansion feature-finding method as well as a detailed features chart using a set of more specific criteria. We next address the performance issue. Section 8 deals with the retrieval effectiveness of expanded queries and Section 9 discusses the computational efficiency of performing AQE. In Section 10 we discuss a few critical issues that must still be solved for moving AQE beyond its experimental status. Section 11 reviews some research directions, and finally, Section 12 offers some conclusions.

2. DOCUMENT RANKING WITH AQE

Most IR systems including search engines rely, totally or in part, on computing the importance of terms that occur in the query and in the documents to determine their answers. The similarity $sim(q, d)$ between query q and document d can be usually expressed as

$$sim(q, d) = \sum_{t \in q \cap d} w_{t,q} \cdot w_{t,d}, \quad (1)$$

where $w_{t,q}$ and $w_{t,d}$ are the weights of term t in query q and document d, respectively, according to the system's *weighting function*. The weight of a term is typically proportional to the term frequency and inversely proportional to the frequency and length of the documents containing the term. This broad formulation accounts for several widely used ranking models that which can be directly or indirectly traced back to it, including vector space model [Salton and McGill 1983], probabilistic relevance model [Robertson et al. 1998], statistical language modeling [Zhai and Lafferty 2001b], and deviation from randomness [Amati et al. 2001].

The ranking scheme of formula 1 can be easily modified to accommodate query expansion, abstracting away from the specific underlying weighting model. The basic input to AQE consists of the original query q and a source of data from which to compute and weight the expansion terms. The output of AQE is a query q' formed by an expanded set of terms with their associated weights w'. The new weighted query terms are used to compute the similarity between query q' and document d

$$sim(q', d) = \sum_{t \in q' \cap d} w'_{t,q'} \cdot w_{t,d}. \quad (2)$$

The most typical data source for generating new terms is the collection itself being searched and the simplest way of weighting the query expansion terms is to use just the weighting function used by the ranking system. If more complex features than single terms are used for query expansion (e.g., phrases), the underlying ranking system must be able to handle such features.

3. WHY AND WHEN AQE WORKS

In most document ranking systems the query terms are connected by an implicit OR. Under this assumption, one advantage of query expansion is that there is more chance for a relevant document that does not contain the original query terms to be retrieved, with an obvious increase in recall. For instance, if the query *Al-Qaeda* is expanded to *Al-Qaeda al-Qaida al-Qa'ida "Osama bin Laden" "terrorist Sunni organization" "September 11 2001,"* this new query does not only retrieve the documents that contain the original term (*Al-Qaeda*) but also the documents that use different spellings or don't directly name it. This observation has originated most early research in AQE, and such a capacity is still very important for search applications in professional domains (e.g., legal, financial, medical, scientific) where the main goal is to retrieve all documents that are relevant to an issue. Notice that a strict recall improvement can be achieved even when the query terms are strictly ANDed together by default, as with some Web search engines, provided that the expanded query can be submitted to the system by using Boolean operators (e.g., AND of ORs).

The additional terms, however, may cause query drift—the alteration of the focus of a search topic caused by improper expansion Mitra et al. [1998]—thus hurting precision. There may be several reasons for this. When an expansion term is

correlated with a single term of the original query rather than with the entire query it may easily match unrelated concepts. This phenomenon may be more serious if the additional term is a proper noun, as pointed out in Vechtomova and Karamuftuoglu [2004]. It is also possible that the set of candidate expansion terms as a whole is not relevant to the original query. This may happen, for instance, when AQE is based on the top documents retrieved in response to the original query and such documents are mostly not relevant. A further reason for a decrease in precision is that the relevant documents that match just the original query terms may move lower down in the ranking after query expansion, even if the additional terms are relevant to the query concept. For example, if the query "Jennifer Aniston" is expanded with "actress," "movie," and "player," a document about a different actress in which such additional terms are well represented may be assigned a higher score than a document about Jennifer Aniston that does not contain the additional terms [Carmel et al. 2002]. That query expansion may result in a loss of precision has been confirmed in some earlier experimental studies (e.g., Voorhees and Harman [1998]).

On the other hand, the effectiveness of IR systems is usually evaluated taking into account both recall and precision. Using a combined recall/precision measure, the overwhelming majority of recent experimental studies agree that AQE results in better retrieval effectiveness, with improvements of the order of 10% and larger (e.g., Mitra et al. [1998], Carpineto et al. [2002], Liu et al. [2004], Lee et al. [2008]). Such findings are important to support the claim that AQE is an effective technique, but this may be not sufficient for the cases when we are primarily interested in precision. However, as explained in the following, several recent studies have pointed out that AQE does not necessarily hurt precision.

One common problem affecting the precision of document ranking is that retrieved documents can often match a query term out of context with its relationships to the other terms. There may be several types of out-of-context matches causing false drops. In Bodoff and Kambil [1998], for instance, five types were identified: polysemy, ordered relationships among terms (e.g., "wars due to crises" versus "crises due to wars"), out of phrase terms (when a query or document phrase is not treated as a single unit), secondary topic keyword (e.g., "siamese cats" versus "cats"), and noncategorical terms (e.g., "tiger" is simultaneously an instance of "mammal" and of "operating system").

The problem of improper partial matching between query and document can be ameliorated by using AQE, to the extent that the additional terms favor a more univocal interpretation of the original query. For example, if the query "tiger, operating system" is expanded with "Mac OS X," the score of the documents about the computer meaning of "tiger" will increase while the score of the documents about different meanings of "tiger" or different operating systems will decrease. This is an example of out-of-phrase term-matching. A similar argument can be applied to the other types of out-of-context matches. Indeed, some recent studies have confirmed that AQE may also improve precision by implicitly disambiguating query terms (e.g., Bai et al. [2005], Carmel et al. [2002], Navigli and Velardi [2003]). In Section 6.2.3 we give an example of this behavior in a situation of practical interest, while the use of word sense disambiguation techniques in IR is discussed in Section 5.3.

Sometimes, AQE achieves better precision in the sense that it has the effect of moving the results toward the most popular or representative meaning of the query in the collection at hand and away from other meanings; e.g., when the features used for AQE are extracted from Web pages [Cui et al. 2003], or when the general concept terms in a query are substituted by a set of specific concept terms present in the corpus that co-occur with the query concept [Chu et al. 2002]. AQE is also useful for improving precision when it is required that several aspects (or dimensions) of a query must be present at once in a relevant document. This is another facet of query disambiguation,

in which query expansion can enhance those aspects that are underrepresented in the original user query [Arguello et al. 2008; Crabtree et al. 2007].

We should emphasize that AQE may not be suitable for all user queries, especially when searching the Web. It has been observed Broder [2002] that most Web queries fall into one of three basic categories: informational, navigational, or transactional. The informational queries (in which the user has a particular information need to satisfy) seem the most relevant to AQE because the user often does not know exactly what he is looking for and and/or he is not able to clearly describe it in words. By contrast, in navigational queries (where the user has a particular URL to find) and transactional queries (where the user is interested in some Web-mediated activity), usually the sought pages are characterized by very specific words that are known to the user.

4. APPLICATIONS OF AQE

Although in this survey we mainly focus on the use of query expansion for improving document ranking, there are other retrieval tasks that may benefit from this technique. We now briefly discuss four areas in addition to document ranking, where the use of AQE has been rather intensive, and then provide pointers to further, more recent, applications.

4.1 Question Answering

The goal of question answering (QA) is to provide concise responses (instead of full documents) to certain types of natural language questions such as "How many kings were there in ancient Rome?". Similar to document ranking, QA is faced by a fundamental problem of mismatch between question and answer vocabularies.

To improve the early document retrieval stage of a QA system, one common strategy is to expand the original question with terms that are expected to appear in documents containing answers to it, often extracted from FAQ data [Agichtein et al. 2004; Harabagiu and Lacatusu 2004]. A recent example in this research line is Riezler et al. [2007], in which the FAQ data are processed by statistical machine translation techniques, as if questions and answers in the corpus were two distinct languages. In this case, the goal of question-answer translation is to learn associations between question words and synonymous answer words. Different approaches to AQE for QA include using lexical ontologies such as WordNet [Harabagiu et al. 2001], shared dependency parse trees between the query and the candidate answers [Sun et al. 2006], and semantic parsing of questions based on roles [Schlaefer et al. 2007], among others.

In the Multilingual Question Answering Track run at the Cross Language Evaluation Forum (CLEF)[3], 2009, three variants of the classical QA task were explored: geographical QA, QA in speech transcripts, and passage retrieval from legal texts. Some authors made use of AQE techniques based on lexical or geographical ontologies, with good [Agirre et al. 2009] or mixed [Flemmings et al. 2009] results.

4.2 Multimedia Information Retrieval

With the proliferation of digital media and libraries, search of multimedia documents (e.g., speech, image, video) has become increasingly important. Most multimedia IR systems perform text-based search over media metadata such as annotations, captions, and surrounding html/xml descriptions. When the metadata is absent, IR relies on some form of multimedia content analysis, often combined with AQE techniques.

[3]www.clefcampaign.org

For example, in spoken document retrieval, the transcription produced by an automatic speech recognition system can be augmented with related terms prior to query time [Singhal and Pereira 1999]. This form of document expansion is very useful for spoken document retrieval since automatic speech transcriptions often contain mistakes, while for plain document retrieval its benefits are more limited [Billerbeck and Zobel 2005; Wei and Croft 2007]. In image retrieval, a typical approach consists of using query examples with visual features such as colors, textures, and shapes, and iteratively refining the visual query through relevance feedback Kherfi et al. [2004]. In video retrieval, both the documents and the queries are usually multimodal, in that they have textual as well as visual aspects. An expanded text query is typically compared against the textual description of the visual concepts and any matched concepts are used for visual refinement. Also, AQE can be directly applied to visual examples represented by low-level feature vectors using relevance or pseudo-relevance feedback (assuming that the top retrieved images are relevant). A review of existing AQE approaches to video retrieval is given in Natsev et al. [2007]. The authors also present an interesting method based on identifying global correlations (not related to a specific query) between terms from the speech transcript and visual concepts; such visual concepts are then used for query expansion.

4.3 Information Filtering

Information filtering (IF) is the process of monitoring a stream of documents and selecting those that are relevant to the user. The documents arrive continuously and the user's information needs evolve over time. Some examples of filtering application domains are electronic news, blogs, e-commerce, and e-mail (see Hanani et al. [2004] for a review). There are two main approaches, collaborative IF (based on the preferences of like-minded users) and content-based IF. The latter technique bears a strong conceptual similarity to IR because the user profile can be modeled as a query and the data stream as a collection of documents [Belkin and Croft 1992].

Better profiles (queries) can be learned using relevance feedback techniques [Allan 1996], or other forms of query expansion, such as based on similar users [Palleti et al. 2007] or on links and anchor textin Wikipedia [Arguello et al. 2008]. In Zimmer et al. [2008], keyword correlation is used to improve the recall in approximate IF—a scenario in which the system is responsible for selecting the best information sources to which a subscription (query) should be submitted.

4.4 Cross-Language Information Retrieval

Cross-language information retrieval (CLIR) deals with retrieving documents written in a language other than the language of the user's query. There has been an increasing interest in CLIR in the last years, thanks to the annual evaluation campaigns run by CLEF and TREC. The traditional approach to CLIR consists of query translation followed by monolingual retrieval, where query translation is performed with machine readable bilingual dictionaries, parallel corpora or machine translation [Koehn 2010]. Regardless of the type of translation resource used, there are usually limitations due to insufficient coverage, untranslatable terms, and translation ambiguity between the source and target languages [Pirkola et al. 2001]. To combat the errors induced by translation, one well known technique is to use query expansion [Ballesteros and Croft 1997]; even when the translation contains no error, the use of semantically similar terms yields better results than those obtainable by literal translation terms alone [Kraaij et al. 2003]. Query expansion can be applied before or after translation, or even at both times; pretranslation yields better results than posttranslation, with a combination being the most effective [Ballesteros and Croft 1998; McNamee and

Mayfield 2002]. A more recent work [Cao et al. 2007] integrates both translation relations and monolingual relations such as term co-occurrence into a unique directed graph in which query translation is performed as a random walk.

4.5 Other Applications of AQE

Other recent applications of AQE include text categorization [Zelikovitz and Hirsh 2000; Hidalgo et al. 2005], search of hidden Web content that is not indexed by standard search engines [Graupmann et al. 2005], query completion on mobile devices [Kamvar and Baluja 2007], training corpora acquisition [Huang et al. 2005], e-commerce [Chen et al. 2004; Perugini and Ramakrishnan 2006], mobile search [Church and Smyth 2007], expert finding [Macdonald and Ounis 2007], slot-based document retrieval [Suryanto et al. 2007], federated search [Shokouhi et al. 2009], and paid search advertising [Broder et al. 2009; Wang et al. 2009].

5. RELATED TECHNIQUES

The word mismatch between query and documents is a long-standing issue in the field of IR. In this section, AQE is put in context with respect to alternative strategies to the vocabulary problem.

5.1 Interactive Query Refinement

There is a vast related literature on interactive query expansion (IQE) and refinement (e.g., Efthimiadis [1996], Baeza-Yates and Ribeiro-Neto [1999]). Its main difference from automatic methods is that the system provides several suggestions for query (re)formulation, but the decision is made by the user. From a computational point of view, IQE and AQE share the first two computational steps, namely data acquisition and candidate feature generation, whereas IQE does not address the subsequent problems of feature selection and query reformulation.

One of the best known systems of this kind is Google Suggest, which offers real-time hints to complete a search query as the user types. IQE has the potential for producing better results than AQE Kanaan et al. [2008], but this generally requires expertise on the part of the user Ruthven [2003]. From a usability point of view, IQE gives the user more control over the query processing, which is a aspect lacking in AQE (see Section 10.3). Although in this article we focus on fully automatic methods for single-query searches, we do include some innovative techniques mainly developed for term suggestion, which are susceptible to also being used for AQE.

5.2 Relevance Feedback

Relevance feedback takes the results that are initially returned from a given query and uses information provided by the user about whether or not those results are relevant to perform a new query. The content of the assessed documents is used to adjust the weights of terms in the original query and/or to add words to the query. Relevance feedback is often implemented using variants of the Rocchio algorithm [Rocchio 1971], discussed in the following, or the F_4 probabilistic reweighting formulas [Robertson and Sparck Jones 1976; Robertson 1986; Robertson and Walker 2000]. Relevance feedback is covered in several books (e.g., Harman [1992], Baeza-Yates and Ribeiro-Neto [1999], Manning et al. [2008]) and surveys Ruthven and Lalmas [2003]. A dedicated track (the Relevance Feedback track) was run at TREC in 2008 and 2009.

Relevance feedback essentially reinforces the system's original decision, by making the expanded query more similar to the retrieved relevant documents, whereas AQE tries to form a better match with the user's underlying intentions. The specific data source from which the expansion features are generated using relevance feedback may

be more reliable than the sources generally used by AQE, but the user must assess the relevance of the documents. On the other hand, relevance feedback has directly inspired one of the most popular AQE techniques, namely pseudo-relevance feedback (discussed in Section 6.2.3), and it has also provided foundational work for modeling query reformulation in a variety of AQE approaches (see Section 6.4).

5.3 Word Sense Disambiguation in IR

Word sense disambiguation (WSD) is the ability to identify the meaning of words in context in a computationalmanner [Navigli 2009]. WSD is a natural and well known approach to the vocabulary problem in IR [Krovetz and Croft 1992; Lesk 1988; Sanderson 2000]. Early work focused on representing words by the text of their dictionary definitions, or by their WordNet synsets (discussed in Section 6.2.1).[4] However, several experiments suggested that a straightforward application of this technique may not be effective for IR Voorhees [1993], at least as long as the selection of the correct sense definition (or synset) is flawed; e.g., if the precision is no greater than 75%, according to Sanderson [1994]. The work on using WordNet for AQE has continued using more sophisticated methods, described below in the paper.

Rather than relying on short, predefined lists of senses, it may be more convenient to use a corpus as evidence to perform word sense induction. In Schütze and Pedersen [1995], the context of every occurrence of a word is found and similar contexts are clustered to determine the word senses (or word uses). With a correct disambiguation rate of 90%, this paper was the first to show that WSD can work successfully with an IR system, reporting a 7 to 14% improvement in retrieval effectiveness. Given its reliance on corpus analysis, this approach is similar in spirit, to the global AQE techniques discussed in Section 7.2. Another corpus-based WSD technique is described in Véronis [2004]. By applying the metaphor of *small worlds* to word co-occurrence graphs, this technique is capable of discovering low-frequency senses (as low as 1%).

On the whole, however, the application of WSD to IR presents both computational and effectiveness limitations. Mixed evidence has also been reported in a recent series of experiments performed at CLEF 2008 and CLEF 2009, in the Robust–WSD task [Agirre et al. 2009]. Furthermore, a typical query context, as in Web searches, may be too short for sense disambiguation.

5.4 Search Results Clustering

Search results clustering (SRC) organizes search results by topic, thus allowing, in principle, direct access to the documents pertaining to distinct aspects of the given query. In contrast to conventional clustering, SRC algorithms try to optimize not only the clustering structure, but also the quality of cluster labels, because a cluster with a poor description is very likely to be entirelyomitted by the user even if it points to a group of strongly related and relevant documents. Some examples of description-centric SRC algorithms are Clusty[5], Lingo [Osiński and Weiss 2005], and KeySRC [Bernardini et al. 2009], all available for testing on the Internet. A recent review of this relatively large body of literature is given in Carpineto et al. [2009].

The cluster labels produced by SRC algorithms can be naturally seen as refinements of the given query, although they have been typically employed for browsing through the search results rather than for reformulating the query. An explicit link between

[4]Term co-occurence representations are typically used in the computational linguistic community to express the semantics of a term. A comparison with document occurrence representations, more common in IR, is made in Lavelli et al. [2004].
[5]http://clusty.com

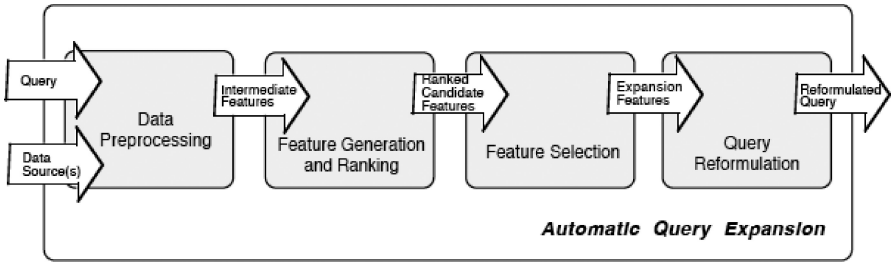

Fig. 1. Main steps of automatic query expansion.

SRC and AQE is made in Kurl and et al. [2005], where the clusters built from top retrieved documents are used as pseudo-queries representing different facets of the original query. This approach can be iterated, although caution must be taken (e.g., by rescoring the documents retrieved at each round) to avoid query drift.

5.5 Other Related Techniques

Other techniques related to AQE include Boolean term decomposition [Wong et al. 1987], spreading activation networks [Crestani 1997], concept lattice-based IR [Carpineto and Romano 2004], random indexing [Sahlgren 2005], and contextual document ranking modeled as basis vectors [Melucci 2008]. Although these methods do not strictly perform query expansion, they have the ability to retrieve documents that do not contain the original query terms, based on particular content relationships among all the terms contained in the collection. Another relevant technique is latent semantic indexing (LSI), which replaces the observed features of documents with a new (smaller) set of uncorrelated features using the singular value decomposition of the term-document matrix [Deerwester et al. 1990]. The relationships between LSI and Rocchio relevance feedback have been theoretically investigated in Efron [2008]. Rocchio is optimal for discriminating between relevant and nonrelevant documents (viewing IR as classification), whereas LSI is optimal for estimating the degree of relevance of a particular document (viewing IR as regression), because projection onto a low-dimension space reduces model variance. Features generated by LSI have been directly used for AQE in Park and Ramamohanarao [2007].

6. HOW AQE WORKS

AQE can be broken down into the four steps shown in Figure 1: preprocessing of data source, generation and ranking of candidate expansion features, selection of expansion features, query reformulation. Each step is discussed, in turn, in the following sections.

6.1 Preprocessing of Data Source

This step transforms the raw data source used for expanding the user query into a format that will be more effectively processed by subsequent steps. It usually consists of a phase of extraction of intermediate features, followed by the construction of appropriate data structures for easy access to and manipulation of such features. Preprocessing of a data source is usually independent of the particular user query that is to be expanded but it is specific to the type of data source and expansion method being considered. The most common preprocessing procedures are discussed in the following.

Many query expansion techniques are based on the information contained in the top-ranked items retrieved in response to the original user query from a collection of

documents. In order to compute the initial retrieval run, it is necessary to index the collection and run the query against the collection index. Indexing usually comprises

(1) text extraction from documents like HTML, PDF, MS Word, and so on (if the collection is made of such documents);
(2) tokenization (extraction of individual words, ignoring punctuation and case);
(3) stop word removal (removal of common words such as articles and prepositions);
(4) word stemming (reduction of inflected or derivational words to their root form);
(5) word weighting (assignment of a score that reflects the importance of the word, usually in each document).

To illustrate, consider the following short HTML fragment.

'Automatic query expansion expands queries automatically.'

The indexed representation, using Porter's stemmer [Porter 1997], and assuming that the weight of a word is simply given by its frequency in the text, is:

automat 0.33, *queri* 0.33, *expan* 0.16, *expand* 0.16.

As a result, each document is represented as a set of weighted terms, with a complementary inverted index file, which maps terms to documents at query time. The indexing system may also store term positions, to provide proximity-based search. When the collection used for query expansion is the same as the one being searched (e.g., Attar and Fraenkel [1977], Xu and Croft [1996], Robertson et al. [1998], Carpineto et al. [2001], Bai et al. [2005]), the ranking system to which the expanded query will be submitted is typically used to also perform a first-pass ranking. If an external corpus is employed (e.g., Web data for intranet searches, or personal desktop data for Web searches), as in Xu and Croft [2000], Voorhees [2004], Diaz and Metzler [2006], and Chirita et al. [2007], a different IR system will, in general, be necessary; several options are available, such as installing and running a desktop search engine (commercial or freely available), using Web retrieval APIs, or even developing one's own system for document indexing and ranking.

Other AQE techniques, based on corpus analysis, require the extraction of particular features from the collection at hand. These are usually different from those discussed in the preceding, which are employed for indexing purposes by a conventional IR system. A well known approach is Qiu and Frei [1993], where each term is represented as a weighted document vector using nonstandard collection statistics. Another example is Crouch and Yang [1992], which builds a statistical thesaurus by first clustering the whole document collection via the complete link clustering algorithm.

Some query expansion techniques require preprocessing procedures tailored to certain data sources. For example, if query expansion makes use of anchor texts, one needs to parse a hyperlinked collection to extract the text content of anchor tags, and to further process such texts to normalize them and/or remove those that contain too few or too many terms [Kraft and Zien 2004]. Clickthrough records (query, URL) extracted from search engine logs are another source of data for query expansion (e.g., Cui et al. [2003], Billerbeck et al. [2003]). In this case, besides extracting from the user logs, the sequence of characters comprising the query and the corresponding documents clicked on, it may be useful to remove objectionable content and also to perform some form of query and URL canonicalization to find semantically equivalent strings [Beeferman and Berger 2000].

In the approaches discussed so far, preprocessing is applied to a given data source. This is the predominant situation, but there are exceptions. The data source may be selected from multiple choices, as in Gauch et al. [1999] and He and Ounis [2007], or

even built from scratch. Two examples of the latter are Riezler et al. [2007] and Bai et al. [2007]. In Riezler et al. [2007], a collection of FAQs is automatically built by first using Web queries such as "inurl:faq" and subsequently applying machine learning techniques to extract the actual FAQs from the retrieved set of pages. In Bai et al. [2007], several strategies for constructing domain models (topic profiles) to which the queries will be assigned for expansion are tested. Such strategies involve the utilization of the documents contained in the Open Directory Project[6], or the top Web answers to user-defined topics. Offline Web-based construction of term vectors representing fixed topics is also performed in Finkelstein et al. [2002]. In all these cases, an earlier preprocessing procedure is necessary to acquire the source data in the first place, prior to the strict data preprocessing step dealt with in this section.

6.2 Generation and Ranking of Candidate Expansion Features

In the second stage of AQE, the system generates and ranks the candidate expansion features. The reason that feature ranking is important is that most query expansion methods will only choose a small proportion of the candidate expansion features to add to the query.

The input to this stage is the original query and the transformed data source; the output is a set of expansion features, usually with associated scores. The original query may be preprocessed to remove common words and/or extract important terms to be expanded (the importance being approximated e.g., by their inverse document frequency).

We classify the techniques used to execute candidate generation and ranking according to the type of relationship between the expansion features generated and the query terms (after query preprocessing, if any).

6.2.1 One-to-One Associations. The simplest form of candidate generation and ranking is based on one-to-one associations between expansion features and query terms, i.e., each expansion feature is related to a single query term. In practice, one or more expansion features are generated and scored for each query term, using a variety of techniques.

One of the most natural approaches is to rely on linguistic associations, such as using a stemming algorithm to reduce different words to the same stem. A stemmer may remove inflected forms of a word that strictly follow the language syntax (e.g., singular/plural of nouns, tenses of verbs), or it may also remove derivational forms. In the latter case, the stem will not, in general, coincide with the morphological root of the word. For instance, using Porter's derivational stemming algorithm [Porter 1997], the words "generalizations," "generalization," "generalize," and "general" would be reduced to the same stem: "gener." Clearly, the latter approach is more powerful but it is prone to errors due to overgeneralization.

Another common linguistic technique is to find synonyms and related words of a query word from a thesaurus, most usually from WordNet (e.g., Voorhees [1994], Mandala et al. [1998]). The WordNet lexicon [Miller et al. 1990], available online[7], groups English words into sets of synonyms called synsets and records various lexical semantic relations between these synonym sets. In particular, it includes hypernym/hyponym relationships among noun synsets that can be interpreted as generalization/specialization relations between the concepts corresponding to such

[6]http://dmoz.org
[7]http://wordnet.princeton.edu/

synsets. For instance, there are three synsets with the noun "spider" in WordNet, each with a specific sense—for zoology, computer science, and cooking. The synset with the computer science meaning is *spider, wanderer*, which is defined as "a computer program that prowls the internet looking for..." and has one direct hypernym ({*program*, *programme*, *computer program*, *computer programme*}) and no hyponyms.

Expansion feature generation from WordNet requires selecting one synset for a given query term, thus solving the ambiguity problem, and then traversing the hierarchy by following its typed links. In order to choose a synset with a similar meaning to the query term, the adjacent query terms can be best matched with the concepts present in each synset containing the query term. After selecting the most relevant synset, one might consider for query expansion, all the synonyms of the query term in the synset plus the concepts contained in any synset directly related to it, usually with different weights (see Section 6.4). Using this approach on the query "spider program" for instance, it would first select the WordNet node with the computer meaning of spider, and then the following candidate query expansion features would be generated: "wanderer," "programme," "computer program," "computer programme."

A radical departure from the linguistic approach consists of generating associations automatically by computing term-to-term similarities in a collection of documents. The general idea is that two terms are semantically related if they appear in the same documents, just as two documents are considered similar if they contain the same terms. Two simple measures of similarity are the Dice coefficient and the Jaccard index. Given terms terms u and v, the Dice coefficient (D) is defined as

$$D = \frac{2 \cdot df_{u \wedge v}}{df_u + df_v}, \tag{3}$$

where $df_{u \wedge v}$ is the number of documents that contain both u and v, and df_u, df_v are the number of documents that contain u and v, respectively.

The Jaccard index (J) is defined as

$$J = \frac{df_{u \wedge v}}{df_{u \vee v}}, \tag{4}$$

where $df_{u \vee v}$ is the number of documents that contain u or v.[8]

A more general approach is the following. Consider a term-document matrix A, where each cell $A_{t,d}$ is a weight $w_{t,d}$ for term t and document d. If we calculate $C = AA^T$, then C is a term-term correlation matrix, where each element $c_{u,v}$ is a correlation (similarity) score between terms u and v given by

$$c_{u,v} = \sum_{d_j} w_{u,j} \cdot w_{v,j}. \tag{5}$$

Using this formula, we can compute the correlation between each term of the query and each term in the document collection. To take into account the relative frequency of terms, it is preferable to generate normalized correlation factors, e.g. by the cosine similarity: $\frac{c_{u,v}}{\sqrt{\sum_{d_j} w_{u,u}^2 \cdot \sum_{d_j} w_{v,v}^2}}$.

Depending on how the set of documents and the weighting function are chosen, Formula (5) can give rise to conceptually different term-to-term correlation methods. One well known technique, first proposed in Attar and Fraenkel [1977], relies on the set of documents returned in response to the original query and makes use of term

[8]The Dice coefficient and the Jaccard index are related: $D = 2J/(1 + J)$ and $J = D/(2 - D)$.

frequency to weight the terms. We will see more elaborated techniques that can be traced back to Formula (5) in Section 6.2.2.

Computing co-occurrence of terms in the whole document is simple but it has the disadvantage that position is not taken into account, whereas two terms that co-occur in the same sentence seem more correlated than two terms that occur distantly within a document. This aspect is usually addressed by considering term proximity; using only restricted textual contexts such as windows of fixed length for measuring co-occurrence of terms. However, the simple co-occurrence, whether in a large or small context, does not necessarily mean that the terms are correlated. For instance, the word "verdi" is correlated with the word "giuseppe" in a music book, whereas the same correlation will not hold for a telephone book, because in the latter case the surname "verdi" cooccurs with many names other than "giuseppe."

A more comprehensive measure for word association that incorporates term dependency is mutual information [Church and Hanks 1990; van Rijsbergen 1979], defined as

$$I_{u,v} = \log_2 \left[\frac{P(u,v)}{P(u) \cdot P(v)} + 1 \right], \tag{6}$$

where $P(u, v)$ is the joint probability that u and v co-occur within a certain context, usually a window of interest, and $P(u)$ and $P(v)$ are the probability of occurrence of terms u and v, respectively. Such probabilities can be estimated, for instance, by relative frequency counts.

Notice that the mutual information is symmetric: $I(u, v) = I(v, u)$. As word order matters (e.g., compare "word processing" to "processing word"), it is preferable to consider an asymmetric version, in which $P(u, v)$ is the probability that v strictly follows u. The mutual information is zero if there is a zero co-occurrence, equal to one if u and v are independent, and equal to $log_2(\frac{1}{P(u)} + 1)$ if v is perfectly associated with u. One of its disadvantages is that it tends to favor rare terms over common terms, because $I(u, v)$ will increase if $P(v|u)$ is fixed, but $P(u)$ decreases. This problem may become more acute for sparse data, which is most relevant to us.

Alternatively, we could consider the classical definition of conditional probability to measure the strength of the association of term v to term u

$$P(v|u) = \frac{P(u,v)}{P(u)}. \tag{7}$$

The conditional probability can be computed by dividing the number of contexts (e.g., phrases) in which terms u and v co-occur by the number of contexts in which term u occurs. This popular approach (e.g., Schütze and Pedersen [1997], Bai et al. [2005]) is similar to the definition of confidence of association rules in data mining problems [Agrawal et al. 1993]. In fact, association rules have been explicitly used for finding expansion features correlated with the query terms [Latiri et al. 2004; Song et al. 2007]. One disadvantage of this approach is that associations with high confidence may hold by chance (e.g., when the two terms are statistically independent).

Expansion features can also be generated by mining user query logs, with the goal of associating the terms of the original query with terms in past related queries. As the texts extracted from such data (possibly after preprocessing—see Section 6.1) are usually very short, the standard correlation techniques based on term frequency cannot be applied. In fact, several additional contextual clues extracted from the query logs have been used to help identify useful associations, such as considering queries that occurred in the same session (e.g., successive queries issued by a single user [Jones et al. 2006]) or queries that yielded similar sets of presumably relevant documents (e.g., by deploying the bipartite graph induced by queries and user clicks [Beeferman

and Berger 2000]). These latter types of evidence do not depend on the content of queries and documents and are thus especially useful when content-based approaches are not applicable. We will return to this in Section 7.

6.2.2 One-to-Many Associations. One-to-one associations tend to add a term when it is strongly related to one of the query terms. However, this may not accurately reflect the relationships of the expansion term to the query as a whole. This problem has been analyzed in Bai et al. [2007]. For example, while the word "program" may well be highly associated with the word "computer," an automatic expansion of all queries containing "program" with "computer" might work well for some queries (e.g., "Java program," "application program"), but not for others (e.g., "TV program," "government program," "space program"). Here again we come across the issue of language ambiguity.

One simple approach to one-to-many associations is to extend the one-to-one association techniques described in the previous section to the other terms in the query. The idea is that if an expansion feature is correlated to several individual query terms, then it is correlated to the query as a whole. In Voorhees [1994], for instance, it is required that a new term extracted from WordNet be related to at least two original query terms before it is included in the expanded query. If we use term-to-term correlations, we might compute the correlation factors of a given candidate expansion term v to every query term, and then combine the found scores to find the correlation to the global query q, e.g. by

$$c_{q,v} = \frac{1}{|q|} \sum_{u \in q} c_{u,v}. \tag{8}$$

A similar approach was suggested in Qiu and Frei [1993] and Xu and Croft [1996], and followed in several other research works [Bai et al. 2005; Cui et al. 2003; Hu et al. 2006; Sun et al. 2006]. The two former papers are interesting not only because they extend the one-to-one correlation paradigm to the whole query, but also because of their particular weighting functions and expansion feature types.

In Qiu and Frei [1993], Formula (5) is used to find term-term correlations in the whole collection, seen as a concept-term space, where documents are used to index terms. The weight of a term in a document is expressed as the product of the frequency of the term in the document by the inverse term frequency associated with that document. The *inverse term frequency* for document d_j is given by $log\frac{T}{DT_j}$, where T is the number of terms in the collection and DT_j is the number of distinct terms in the document d_j. This concept is analogous to the inverse document frequency used for document ranking.

In Xu and Croft [1996], concepts rather than single terms are generated as expansion features. A concept is a group of adjacent nouns in the top retrieved documents; candidate concepts are analyzed using passages (a text window of fixed size) instead of full documents. Formula (5) is applied to compute a term-concept correlation (rather than a term-term correlation), where $w_{u,j}$ is the frequency of the query term u in the j-th passage and $w_{v,j}$ is the frequency of the concept v in the j-th passage. The exact term-concept correlation value is determined by taking into account the inverse frequency of the term and the concept in the passages contained in the whole collection. The correlation factors of each single query term to a given concept are then combined through a function of their product. This method is called *local context analysis*.

The extended one-to-one associations approach can be useful to filter out expansion features that are weakly related to some query terms, but it does not guarantee that an expansion feature that is strongly connected to only one term will be discarded. For

example, if the association of "computer" with "program" is strong enough, "computer" may remain as an expansion term even for the queries "TV program" or "government program."

This problem can be alleviated by adding context words to a term-to-term association that specify under which conditions the association is valid. Such context words, for instance, can be derived as logical consequences from a knowledge base [Lau et al. 2004], or they can be extracted from a corpus using term co-occurrences [Bai et al. 2006]. Considering our example again, if we require that "program" appears with "application" (or "Java"), then we limit the applicability of the association "program"-"computer" to the appropriate contexts.

When query expansion is based on WordNet, the need for relating the expansion features to the entire query, and not to its terms considered in isolation, is even stronger. Voorhees [1994] showed that the latter techniques are usually not effective because they do not guarantee a good word sense disambiguation. This problem, however, can be addressed by analyzing the relationships between the WordNet concepts associated with one query word and the concepts associated with the other (contiguous) query words. Consider as an example the query "tropical storm." The sense of "storm" can be unequivocally determined by observing that a hyponym of the synset {storm, violent storm} is "hurricane," whose definition contains the word "tropical." This and other simple heuristic strategies have been used in Liu et al. [2004]. We will discuss more elaborated disambiguation methods based on WordNet concepts in Section 7.1.

Another, perhaps more principled, approach to finding one-to-many associations is based on combining multiple relationships between term pairs through a Markov chain framework [Collins-Thompson and Callan 2005]. For each query, a term network is constructed that contains pairs of words linked by several types of relations, such as synonyms, stems, co-occurrence, together with transition probabilities. Such relations can be generated from various sources; Collins-Thompson and Callan [2005] makes use of WordNet, Krovetz stemmer, an external corpus and top retrieved documents. Then the words with the highest probability of relevance in the stationary distribution of the term network are selected as expansion features, for they best reflect the multiple aspects of the given query. This approach is more robust with respect to data sparsity and it supports complex inferences involving chains of terms. A similar association paradigm, using a spreading activation model [Anderson 1983], has been successfully applied to solve a language game in which the player has to find a word semantically related to a set of given words [Semeraro et al. 2009].

To overcome the limitations inherent in considering relationships between single terms, one can see the query as a phrase and look for phrases that are related to it. Phrases typically provide a richer context and have a smaller degree of ambiguity than their constituent words, although a similarity assessment at the phrase level may not be straightforward. In Riezler et al. [2007], for instance, the best translation phrases, from which the expansion terms are extracted, are learned from training data; in Liu et al. [2008], the criterion for selecting the best phrases, which are directly used as expansion features, is based on a conceptual distance, measured on WordNet, between the query phrase and keyphrases of its search results.

6.2.3 Analysis of Feature Distribution in Top-Ranked Documents. The techniques described in this section do not fit into either of the previous categories, because they do not try to find features directly associated with the terms in the query, whether single or multiple. The idea is to use the first documents retrieved in response to the original query as a more detailed description of the underlying query topic, from which to extract the most important terms to be used as expansion features. In a sense, the expansion features are related to the full meaning of the query because the extracted

Table I. Main Term-Ranking Functions Based on Analysis of Term Distribution in Pseudo-Relevant Documents

Reference	Function	Mathematical form
[Rocchio 1971]	Rocchio's weights	$\sum_{d \in R} w(t,d)$
[Robertson and Sparck Jones 1976]	Binary independence model (BIM)	$log \frac{p(t\|R)\,[1-p(t\|C)]}{p(t\|C)\,[1-p(t\|R)]}$
[Doszkocs 1978]	Chi-square	$\frac{[p(t\|R)-p(t\|C)]^2}{p(t\|C)}$
[Robertson 1990]	Robertson selection value (RSV)	$\sum_{d \in R} w(t,d) \cdot [p(t\|R)-p(t\|C)]$
[Carpineto et al. 2001]	Kullback-Leibler distance (KLD)	$p(t\|R) \cdot log \frac{p(t\|R)}{p(t\|C)}$

terms are those that best characterize the pseudo-relevant documents as a whole, but their association with the query terms is not analyzed explicitly.

A simple approach, inspired by Rocchio's method for relevance feedback [Rocchio 1971], is to assign a score to each term in the top retrieved documents by a weighting function applied to the whole collection. The weights collected by each term are then summed up and the resulting score is used to sort the set of terms. This approach, termed pseudo-relevance feedback (or retrieval feedback, or blind feedback), is simple and computationally efficient, but it has the disadvantage that each term weight may reflect more the usefulness of that term with respect to the entire collection rather than its importance with respect to the user query.

This issue can be addressed by studying the difference in term distribution between the subsets of (pseudo-)relevant documents and the whole collection. It is expected that terms with little informative content will have the same (random) distribution in any subset of the collection, whereas the terms that are most closely related to the query will have a comparatively higher probability of occurrence in the relevant documents. Following this general paradigm, various functions have been proposed that assign high scores to the terms that best discriminate relevant from nonrelevant documents.

In Table I we show some well known term-ranking functions, including Rocchio's weights. The notation is the following: t indicates a term, $w(t,d)$ is the weight of t in pseudo-relevant document d, $p(t|R)$ and $p(t|C)$indicate the probability of occurrence of t in the set of pseudo-relevantdocuments R and in the whole collection C, respectively. The list in Table I is not exhaustive. Other term-scoring functions, including variants of those reported in this article, are considered in Efthimiadis [1993], Carpineto et al. [2001], and Wong et al. [2008].

The estimation of probabilities in the expressions in Table I is an important issue because it might affect performance results. To compute the probability of occurrence of a term t in X (whether the set of pseudo-relevant documents R or the whole collection C), the maximum likelihood criterion is often adopted—the ratio between the number of occurrences of t in X, treated as a long sequence of terms, and the number of terms in X. A different probability estimate is to use the fraction of documents in X that contain the term t. This latter criterion, generally used to compute the binary independence model (BIM) and Robertson selection value (RSV) functions, has also been applied to Chi-square and Kullback-Leibler distance (KLD) in a recent experimental study [Wong et al. 2008], with very good results.

Each term-ranking function has its ownrationale, and the results produced by their application may be very different. In particular, it has been shown that the ordered sets of expansion terms suggested for each query by the different functions are largely uncorrelated [Carpineto et al. 2002]. However, several experiments suggest that the choice of the ranking function does not have a great impact on the overall system performance as long as it is used just to determine a set of terms to be used in the expanded query [Salton and Buckley 1990; Harman 1992; Carpineto et al. 2001]. By contrast, we will see in Section 6.4 that the scores produced by different functions can make a big difference if they are used not only to select but also to reweight the expansion terms.

6.2.4 Query Language Modeling. Another commonly-used approach to AQE is to build a statistical language model for the query, specifying a probability distribution over terms. The best terms for query expansion are those with the highest probabilities. These techniques are usually referred to as model-based. The two main representatives are the *mixture model* [Zhai and Lafferty 2001a] and the *relevance model* [Lavrenko and Croft 2001], both making use of the top retrieved documents. They are described in the following.

In the former method, similarly to term-ranking functions based on distribution difference analysis, one tries to build a query topic model from the top-ranked documents by extracting the part that is most distinct from the whole document collection. As top-ranked documents are likely to contain both relevant and background (or even irrelevant) information, they can be represented by a mixture generative model that combines the query topic model θ_T (to be estimated) and the collection language model. The log-likelihood of top-ranked documents is as follows, where R is the top-ranked document set, $c(t, d)$ is the number of the occurrences of t in d, and λ is the interpolation weight.

$$log\ p(R|\theta_T) = \sum_{d \in R} \sum_{t} c(t, d)\ log((1 - \lambda)\ p(t|\theta_T) + \lambda\ p(t|C)). \tag{9}$$

The expectation-maximization (EM) algorithm [Dempster et al. 1977] is then used to extract the topic model so as maximize the likelihood of the top-ranked documents (assuming that λ has a non-zero value). Compared to the term-ranking functions illustrated in the preceding, the mixture model has a stronger theoretical basis but there is one parameter (λ) that needs to be set and it may be more difficult to compute.

In the relevance model approach, it is assumed that both the query and the top-ranked documents are samples from an unknown relevance model $theta_{REL}$. To approximate such a model, the probability of term t is related to the conditional probability of observing that term given that we just observed the original query terms. By assuming that the k query terms q_i and the document terms are sampled identically and independently, the following estimate can be derived [Lavrenko and Croft 2001].

$$p(t|\theta_{REL}) = \sum_{d \in R} p(d)\ p(t|d) \prod_{i=1}^{k} p(q_i|d). \tag{10}$$

This model has been widely used recently. As it does not rely on distribution difference analysis, it is more similar in spirit to the Rocchio method. Operationally, its main difference from Rocchiois that top-ranked documents are weighted such that documents further down the list have smaller and smaller influence on word probabilities [Lavrenko and Allan 2006].

An interesting generalization of the relevance model that takes the term dependencies into account is described in Metzler and Croft [2007]. By modeling the relevance

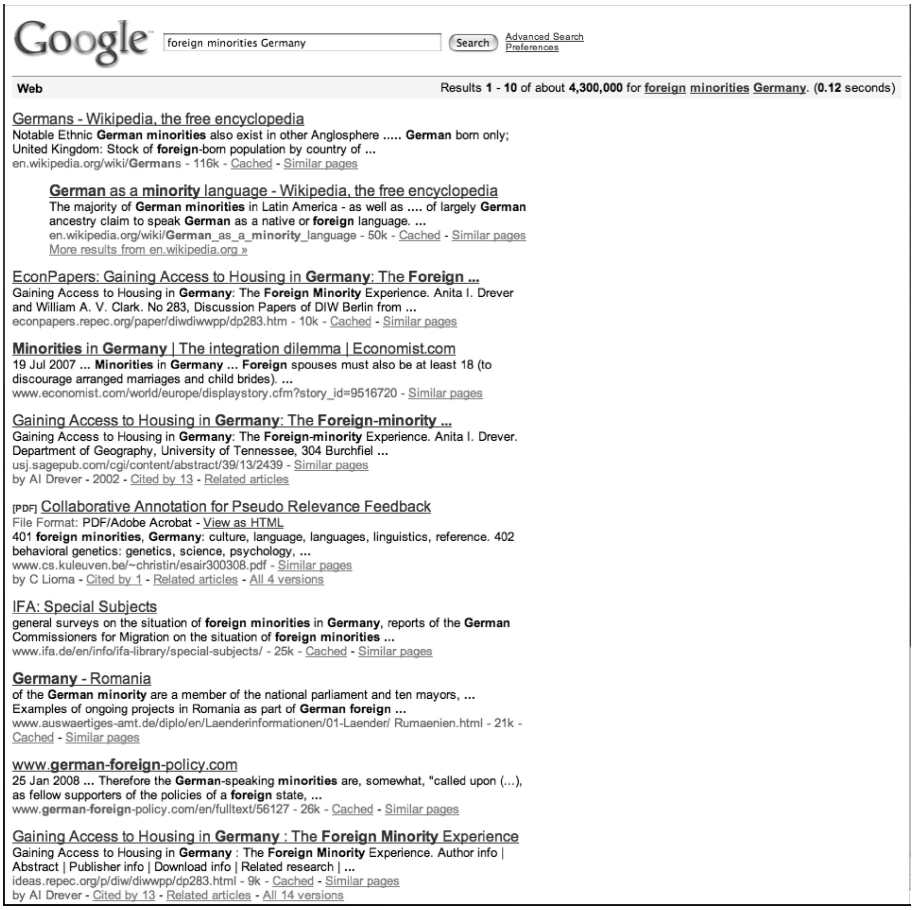

Fig. 2. The first ten results returned by Google in response to the query "foreign minorities Germany" (as of May 2009).

distribution with Markov random fields, a wider set of features is used, that includes not only simple term occurrence statistics but also proximity-based features, such as the number of times the query terms appear ordered or unordered within a window of fixed size. The same method can also generate multi-term expansion concepts, although such concepts were not found to be highly effective, probably due to the correlation between their constituent single terms.

6.2.5 A Web Search Example. To give an impression of the features generated by different expansion methods in a practical application, consider the following example. Suppose you are interested in retrieving Web pages about foreign minorities in Germany. Figure 2 shows the first results page returned by Google in response to the query "foreign minorities Germany" (as of April 2009). Notice that due to improper matching with the query terms, five out of the first ten results are non-relevant to the query (e.g., they are about German minorities living abroad).

To help focus the search, we performed automatic query expansion. Using just the first thirty results (title + snippet) returned by Google as a data source, we applied several expansion-feature generation methods illustrated in the preceding, recapitulated in the first column of Table II. Preprocessing was common to all methods and

Table II. Expansion Features (with Associated Scores) Generated by Several Methods for the Query "Foreign Minorities Germany" by Analyzing the First Thirty Results Returned by Google on the Same Query

Method	Expansion features
Query-term correlation matrix	west (0.415), workers (0.391), policy (0.380), republic (0.326), housing (0.378), access (0.378), language (0.378), cultural (0.378)
Mutual information	integration (4.355), jstor (4.355), reports (4.355), description (4.241) european (0.319), continental (0.319), cultural (0.319), language (0.319)
Local context analysis	housing Germany (26.422), access housing (20.644), minority experience (18.446), foreign policy (16.759), books result (16.586), west germany (15.749), minorities groups (12.718), joschka fischer (10.422)
Rocchio's weights	joschka (1.121), poland (0.865), shareholders (0.726), romania (0.695) danish (0.668), fischer (0.621), frisians (0.618), sorbs (0.580)
Binary independence model	frisians (10.183), sorbs (9.036), joschka (8.878), hillard (6.669) gaining (2.482), shareholders (1.848), fischer (1.304), continental (0.459)
Chi-square	frisians (4.176), sorbs (1.881), joschka (1.685), hillard (0.358) google (0.061), number (0.046), history (0.041), books (0.036)
Robertson selection value	joschka (0.004), gaining (0.002), poland (0.002), frisians (0.002) sorbs (0.002), shareholders (0.001), hillard (0.001), fischer (0.001)
Kullback-Leibler distance	frisians (0.036), sorbs (0.032), joschka (0.032), hillard (0.024), gaining (0.017), poland (0.005), fischer (0.004), clark (0.002)
Relevance model	poland (0.0083449), language (0.0041835), description (0.0041832), european (0.0041820), cultural (0.0041815), continental (0.0041814), west (0.00418107), integration (0.0041806)

consisted of HTML tag stripping, text tokenization, stop wording, and word stemming. The query-term correlation was found by computing the correlations with the single query terms with Formula (5) and then taking their arithmetic mean. A similar procedure was used for the mutual information scores, where we used a window size equal to three, to compute the term-term correlations. The LCA method was approximated considering the snippets as passages and estimating the frequency of the concepts in the Web collection with the frequency counts returned by Google on the candidate concepts (submitted in quotes). The weights in Rocchio and RSV were computed using a simple $tf - idf$ function (proportional to the term frequency in the search result and inversely proportional to the frequency of the Web documents containing the term). For the other methods, we estimated the conditional probabilities $p(t|D)$ as the frequency of term t in document(s) D. For the relevance model, we also used Laplace smoothing to eliminate zeros, and a constant probability value for pseudo-relevant documents. The results produced by each method (expansion features + scores or probabilities) are shown in Table II.

Notice that "frisians" and "sorbs" (two minorities living in Germany) were the first suggestions by BIM, Chi-square, and KLD, and they were also present in the list of expansion terms produced by the other term distribution-based methods. In Figure 3 we see the Google results when the original query was expanded with "frisians" and "sorbs." The difference from the unexpanded case is striking; all the first ten results appear to be relevant to the query, while the overall number of retrieved results reduced from 4,100,000 to 1,610. Judging from Figure 3, the new results cover not only Frisians and Sorbs but also the other minorities. This example shows that it was possible to generate in a very efficient manner, a set of expansion features that produced a more accurate model of the query topic, thus filtering out those pages that spuriously matched the shorter description.

Before concluding this section, we would like to note that the methods for generating query expansion features can themselves take advantage of an earlier query expansion

Fig. 3. The first ten results returned by Google in response to the expanded query "foreign minorities Germany sorbs frisians" (as of May 2009).

step. The idea is to use the query augmented with some context as an input, instead of the mere user query. In Finkelstein et al. [2002], for instance, the authors use the text surrounding the marked query, assuming that search is initiated from a document the user views. This amounts to performing a double query expansion, in which the first expansion is used to reduce query ambiguity and increase the accuracy of the procedure that generates the actual expansion features from the augmented query and the data source. An early stage of query expansion with a similar goal is also used in Jones [1993], to find the WordNet nodes that best match the query terms; such nodes are the starting points to generate the expansion features.

6.3 Selection of Expansion Features

After ranking the candidate features, the top elements are selected for query expansion. The selection is made on an individual basis, without considering the mutual dependencies between the expansion features. This is of course a simplifying assumption, although there are some experimental results that seem to suggest that the independence assumption may be justified [Lin and Murray 2005].

Usually only a limited number of features is selected for expansion, partly because the resulting query can be processed more rapidly, partly because the retrieval effectiveness of a small set of good terms is not necessarily less successful than adding all candidate expansion terms, due to noise reduction (e.g., Salton and Buckley [1990], Harman [1992]).

Some research has been carried out on the optimum number of features to include and thereare differing suggestions ranging from five–ten features [Amati 2003; Chang et al. 2006] to a few hundred [Bernardini and Carpineto 2008; Buckley et al. 1995; Wong et al. 2008]. On the other hand, the performance decrease associated with non-optimal values is usually modest Carpineto et al. [2001], and most experimental studies agree that the number of expansion features is of low relevance. The typical choice is to use 10–30 features. When the feature scores can be interpreted as probabilities, one can select only the terms having a probability greater than a certain threshold; e.g., $p = 0.001$, as in Zhai and Lafferty [2001a].

Rather than concentrating on finding an optimal number of expansion terms, it may be more convenient to adopt more informed selection policies. It has been shown that different queries have a varying optimal number of expansion features [Billerbeck and Zobel 2004a; Buckley and Harman 2003; Cao et al. 2008], and that many expansion terms—about one third in Cao et al. [2008]—are harmful to retrieval performance. In fact, if one were able to select exactly the best features for each query, the performance improvement would be much higher than usually achieved [Carpineto et al. 2002; Cao et al. 2008].

To go beyond a straightforward selection based on the ranks assigned to candidate features, several methods that employ additional information have been proposed. One technique [Carpineto et al. 2002] consists of using multiple term-ranking functions and selecting for each query the most common terms (e.g., based on majority vote). A similar idea is exploited in Collins-Thompson and Callan [2007], with the difference that multiple feedback models are created from the same term-ranking function by resampling documents and by generating variants of the original query. The authors argue that in this way it is possible to remove noise expansion terms as well as focus on expansion terms related to multiple query aspects. Another strategy consists of choosing a variable amount of expansion depending on the query difficulty. In Chirita et al. [2007], the number of expansion terms is a function of the ambiguity of the original query in the Web (or in the personal information repository of the user), as measured by the clarity score [Cronen-Townsend and Croft 2002].

In Cao et al. [2008], the authors use a classifier to discriminate between relevant and irrelevant ranked expansion terms. To learn the Support Vector Machine classifier parameters, a training set is created in which single terms are labeled as good or bad depending on whether they improve or hurt retrieval performance and each term is described by a set of features such as co-occurrence and proximity with query terms. The selection of the best expansion terms for a given query (including zero terms) is explicitly cast as an optimization problem in Collins-Thompson [2009]. By optimizing with respect to*uncertainty sets* defined around the observed data (e.g., using query perturbations and topic-specific constraints such as aspect balance, aspect coverage and support of the query), the system mitigates the risk-reward tradeoff of expansion.

6.4 Query Reformulation

The last step of AQE is query reformulation, namely how to describe the expanded query that will be submitted to the IR system. This usually amounts to assigning a weight to each feature describing the expanded query—termed query

reweighting—but there are other approaches that will be discussed at the end of this section.

The most popular query reweighting technique is modeled after Rocchio's formula for relevance feedback [Rocchio 1971] and its subsequent improvements [Salton and Buckley 1990], adapted to the AQE setting. A general formulation is the following, where q' is the expanded query, q is the original query, λ is a parameter to weight the relative contribution of query terms and expansion terms, and $score_t$ is a weight assigned to expansion term t.

$$w'_{t,q'} = (1 - \lambda) \cdot w_{t,q} + \lambda \cdot score_t \tag{11}$$

When the expansion terms are extracted from pseudo-relevant documents and their score is computed using the documents, or Rocchio's, weights (see the first function in Table I), it is easy to show that the expanded query vector computed by Expression (11) moves towards the centroid of pseudo-relevant documents (according to the document weights). However, the benefits of taking into account the term distribution difference between the pseudo-relevant documents and the whole collection to select the expansion terms may be reduced if we reweight such terms by Rocchio's weights. The rationale is that terms that were correctly ranked higher (because more relevant to the specific query at hand) will be downweighted if their relevance value with respect to the entire collection being searched is low. This observation has been confirmed in several experiments where the use of a distribution difference-based scoring function for both query expansion and reweighting achieved the best retrieval effectiveness, not only for English (e.g., Carpineto et al. [2001], Wong et al. [2008]) but also for other European [Amati et al. 2003] and Asian languages [Savoy 2005]. Even a simple reweighting scheme based on an inverse function of term ranks may produce good results (e.g., Carpineto et al. [2002], Hu et al. [2006]).

Notice that as the document-based weights used for the unexpanded query and thedistribution difference-based scores used for the expansion terms have different scales, their values must be normalized before summing them in Expression (11). Several simple normalization schemes, discussed in Wong et al. [2008], have been proposed; usually they produce comparable results, although more powerful methods that not onlyscale data into the same range but also increase its uniformity could be more effective [Montague and Aslam 2001].

The value of λ in Expression (11) can be adjusted so as to optimize performance, if training data are available. A typical default choice is to give more importance to the original query terms; e.g, twice as much as the expansion terms. Another possibility is to use a parameter-free query reweighting formula such as proposed in Amati [2003]. A more powerful approach is to adaptively determine how much weight one should put on expansion information. In Lv and Zhai [2009], considering a relevance feedback setting, the authors use a learning approach to predict the optimal value of λ for each query and each set of feedback documents, exploring a number of features related to the discrimination of query and documents (such as length, entropy, and clarity) and to the divergence between query and feedback documents.

Formula (11) can also be used when the expansion features have been extracted from a thesaurus or WordNet. The weightings may be based on criteria such as number of connections, number of co-occurrences, path length, and type of relationship [Jones 1995]. In Voorhees [1994], for instance, the expanded query vector is comprised of subvectors of eleven different concept types with an associated importance weight: one for original query terms, one for synonyms, and one each for the other relation types contained within the noun portion of WordNet.

If document ranking is performed through a language modeling approach, the query reweighting step of AQE is naturally supported. In the basic language modeling

framework, the most relevant documents are those that minimize the Kullback-Leibler divergence between the query language model and the document language model:

$$sim(q,d) \propto \sum_{t \in V} p(t|\theta_q) log \frac{p(t|\theta_q)}{p(t|\theta_d)}. \quad (12)$$

In Formula (12), the query model is usually estimated considering only the original query words, while the document model is estimated also taking into account unseen words through probability smoothing, for example, by the Jelinek-Mercer interpolation Jelinek and Mercer [1980]: $p(t|\theta'_d) = (1-\lambda) \cdot p(t|\theta_d) + \lambda \cdot p(t|\theta_C)$. Thus, the question arises as to whether it is possible to create a better query model by finding related words with their associated probabilities and then using the corresponding query expansion model (QEM) to smooth the original query model, in the same way as the document model is smoothed with the collection model. Various methods for creating a query expansion model have been explored, based not only on feedback documents [Lavrenko and Croft 2001; Zhai and Lafferty 2001a], but also on term relations [Bai et al. 2005], and domain hierarchies [Bai et al. 2007]. Regardless of the specific generation method, the final expanded query model (computed with the Jelinek-Mercer interpolation) is given by

$$p(t|\theta'_q) = (1-\lambda) \cdot p(t|\theta_q) + \lambda \cdot p(t|\theta_{QEM}), \quad (13)$$

which can be seen as a generalization of Expression (11).

Query reweighting is common in AQE but it is not always performed. One simple alternative approach is to increase the number of features describing the query without performing query reweighting at all, as in our example in Figure 3. Another approach consists of increasing the number of query features and then applying a modified version of the weighting function used by the ranking system to explicitly deal with the expansion features, in contrast to ranking the documents using the system's basic weighting function in conjunction with a reweighted expanded query. A well known example is Robertson and Walker [2000], used to extend the Okapi BM25 ranking function [Robertson et al. 1998].

In other cases, it is produced by a Boolean query [Graupmann et al. 2005; Liu et al. 2004], or more generally, a structured query [Collins-Thompson and Callan 2005]. In Kekäläinen and Järvelin [1998], it was shown that the probabilistic AND operator, incombination with maximally expanded query aspects, was very effective for query expansion. Nowadays, there are several search query languages that allow specification of general concepts including Boolean filtering, phrase matching, and term proximity, among others. For instance Arguello et al. [2008], using Indri[9], the query "DSLR camera review" can be expressed as:

```
#weight (   0.8   #combine (   DSLR camera review )
            0.1   #combine (   #1 ( DSLR camera )
                               #1 ( camera review )
                               #1 ( DSLR camera review ) )
            0.1   #combine (   #uw8 ( DSLR camera )
                               #uw8 ( camera review )
                               #uw8 ( DSLR review )
                               #uw12 ( DSLR camera review ) ) ),
```

[9]http://www.lemurproject.org/indri/

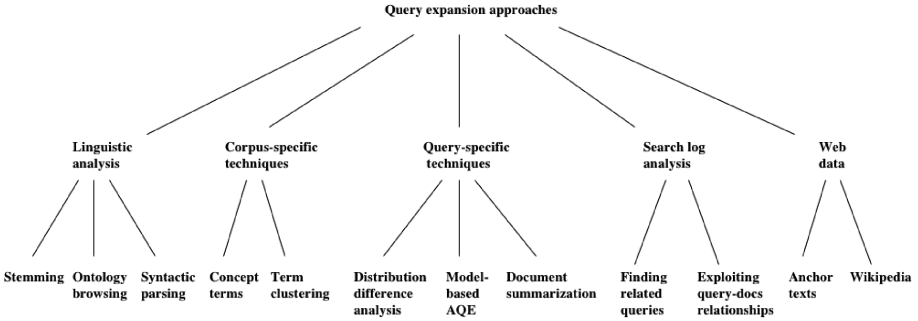

Fig. 4. A taxonomy of approaches to AQE.

where the first line is a unigram query, the second group is a query of exact phrases, and the third group is a query of unordered windows (uw) with a specified size.

7. A CLASSIFICATION OF APPROACHES

AQE techniques can be classified into five main groups according to the conceptual paradigm used for finding the expansion features: linguistic methods, corpus-specific statistical approaches, query-specific statistical approaches, search log analysis, and Web data. Each group can then be further split into a few subclasses, thus yielding the general taxonomy shown in Figure 4. In this section we discuss the main characteristics of the groups in the taxonomy. Then we provide a detailed feature chart, where single AQE techniques can be analyzed and compared to one another along a number of specific dimensions.

7.1 Linguistic Analysis

These techniques leverage global language properties such as morphological, lexical, syntactic and semantic word relationships to expand or reformulate query terms. They are typically based on dictionaries, thesauri, or other similar knowledge representationsources such as WordNet. As the expansion features are usually generated independently of the full query and of the content of the database being searched, they are usually more sensitive to word sense ambiguity.

Using word stems is one of the simplest and earliest language-specific AQE technique. The stemming algorithm can be applied either at indexing time (only the document word stems are stored and then they are matched to the query word stems), as in most systems (e.g., Krovetz [1993], Hull [1996]), or at retrieval time (the original document words are stored and then they are matched to the morphological variants of query terms). The latter strategy may be more effective [Bilotti et al. 2004], but it requires structured querying—an ability that may not be present in alldocument retrieval systems.

Ontology browsing is another well known language-specific AQE technique [Navigli and Velardi 2003]. Knowledge models such as ontologies and thesauri (the distinction between the two is blurred) provide a means for paraphrasing the user's query in context. Both domain-specific and domain-independent ontologies have been used (see Bhogal et al. [2007] for a review of case studies), including the combination of multiple thesauri [Mandala et al. 1999]. Most of the recent work has focused on the use of WordNet. As already remarked, WordNet is very appealing for supporting AQE, but its application may raise several practical issues; e.g., lack of proper nouns and collocations, no exact match between query and concepts, one query term mapping to

several noun synsets. Furthermore, the use of WordNet suffers from the disambiguation problems discussed in Section 5.3. In particular, its use for query expansion is advantageous only if the query words are disambiguated almost exactly ([Voorhees 1994]; Gonzalo et al. [1998]), while word sense disambiguation remains a hard problem [Navigli 2009].

There are several ways to circumvent these difficulties. To increase the coverage of single and multiword concepts, WordNet has been enriched with an automaticallyconstructed thesaurus [Mandala et al. 1998]. The disambiguation issue has been addressed in a more effective manner in some recent papers. In Navigli and Velardi [2005], the authors argue that instead of replacing a given query word with its synonyms, hyperonyms, and hyponyms, it might be better to extract the concepts that pertain to the same semantic domain of query, through other types of definitional information derivable from WordNet, such as gloss words and common nodes. The different types of information present in WordNet can also be combined, e.g., to assign terms in the same query into semantically similar groups, followed by conventional expansion of each group [Gong et al. 2006]. In Liu et al. [2004] and Song et al. [2007], classical Wordnet concepts, extracted by a sequential application of heuristic rules to pairs of query terms, are then integrated with other feature extraction methods.

The third main approach for providing additional linguistic information to the original query is syntactic analysis. The objective is to extract relations between the query terms, which can then be used to identify expansion features that appear in related relations. For example, it is possible to index the user query and the top-ranked snippets by relation paths induced from parse trees, and then learn the most relevant paths to the query [Sun et al. 2006]. The syntactic approach may be most useful for natural language queries; to solve more general search tasks, the linguistic analysis can be more effectively integrated with statistical [Song et al. 2006] or taxonomic information [Liu et al. 2008].

7.2 Corpus-Specific Global Techniques

The techniques in this category analyze the contents of a full database to identify features used in similar ways. Most early statistical approaches to AQE were corpus-specific and generated correlations between pairs of terms by exploiting term co-occurrence, either at the document level, or to better handle topic drift, in more restricted contexts such as paragraphs, sentences, or small neighborhoods. Concept terms [Qiu and Frei 1993] and term clustering [Bast et al. 2007; Crouch and Yang 1992; Schütze and Pedersen 1997] are two classical strategies, already reviewed in the preceding sections. Other approaches to building an association thesaurus are described in Gauch et al. [1999], Hu et al. [2006], Park and Ramamohanarao [2007], and Milne et al. [2007], making use of context vectors, mutual information, latent semantic indexing, and interlinked Wikipedia articles, respectively. This AQE paradigm has also been recently extended with good results to multimedia documents [Natsev et al. 2007]. Note that since global techniques are data-driven, they may not always have a simple linguistic interpretation.

7.3 Query-Specific Local Techniques

Query-specific techniques take advantage of the local context provided by the query. They can be more effective than corpus-specific techniques because the latter might bebased on features that are frequent in the collection but irrelevant for the query athand.

Query-specific techniques typically make use of top-ranked documents. The most commonly used methods are analysis of feature distribution difference and model-based AQE. Both were discussed in depth in the preceding sections.

A different vein of research on query specific-techniques is based on preprocessing top retrieved documents for filtering out irrelevant features prior to the utilization of a term-ranking function. Besides using just Web snippets, several methods for finding more compact and informative document representations have been proposed, such as passage extraction [Xu and Croft 1996] and text summarization [Lam-Adesina and Jones 2001]. In Chang et al. [2006], the document summaries go through a further process of clustering and classification with the aim of finding an even more reduced set of orthogonal features describing each document (termed *query concepts*). In this case, clustering is used to extract intradocument rather than cross-document contextual information, in contrast with the approaches described in Section 5.4.

7.4 Search Log Analysis

The fourth main AQE paradigm is based on analysis of search logs. The idea is to mine query associations that have been implicitly suggested by Web users, thus bypassing the need to generate such associations in the first place by content analysis.

Search logs typically contain user queries, followed by the URLs of Web pages that are clicked by the user in thecorresponding search results page. One advantage of using search logs is that they may encode implicit relevance feedback, as opposed to strict retrieval feedback.

On the other hand, implicit measures are generally thought to be only relatively accurate (see Joachims et al. [2007] for an assessment of the reliability of this assumption) and their effectiveness may not be equally good for all types of users and search tasks [White et al. 2005]. Other problems with their use for AQE are caused by noise, incompleteness, sparseness, and the volatility of Web pages and query [Xue et al. 2004]. Also, the availability of large-scale search logs is an issue.

There are two main AQE techniques based on search logs. The first is to treat the individual queries as documents and extract features from those related to the original user query, with or without making use of their associated retrieval results (e.g., Huang et al. [2003], Jones et al. [2006], Yin et al. [2009]). The second technique, more widely used, consists of exploiting the relation of queries and retrieval results to provide additional or greater context in finding expansion features. Examples of the latter approach include using top results from past queries [Fitzpatrick and Dent 1997], finding queries associated with the same documents [Billerbeck et al. 2003] or user clicks [Beeferman and Berger 2000]), and extracting terms directly from clicked results [Cui et al. 2003; Riezler et al. 2007].

7.5 Web Data

A common Web data source for AQE is represented by anchor texts. Anchor texts and real user search queries are very similar because most anchor textsare succinct descriptions of the destination page. However, in the absence of any implicit user feedback, it is difficult to find the anchor texts that are similar to the query because classical ranking techniques such as Equation (1) do not work well on very short texts. In Kraft and Zien [2004], anchor texts are ranked using several criteria that best relate to the specific nature of the data, such as the number of occurrences of an anchor text (taking into account whether it points to a different site or to the same site) and the number of terms and characters in it. Each anchor text is then assigned a combined rank based on a median aggregation of its individual ranks. At query

time, the highest-ranked anchor texts that have a non-empty intersection with the query are selected as refinement features.

Another interesting method, based on Wikipedia documents and hyperlinks, is proposed in Arguello et al. [2008]. The initial set of candidates associated with a query is restricted by considering only those anchor texts that point to a short set of top-ranked documents from a larger set of top-ranked documents, followed by scoring each candidate proportional to its frequency and inversely proportional to the rank of the documents it links to. Specific categories of Wikipedia articles are used in Xu et al. [2009].

Other types of Web data that can be employed for AQE include FAQs [Riezler et al. 2007] and the Open Directory Project Web pages [Bai et al. 2007].

7.6 A Feature Chart

In Tables III and IV we consider some of the most influential or innovative AQE methods, regardless of their broad conceptual paradigms, and provide a detailed classification along five specific problem dimensions. The methods are ordered chronologically.

8. RETRIEVAL EFFECTIVENESS

The retrieval effectiveness of AQE systems is typically evaluated by executing each querytwice, with and without query expansion, and then comparingthe two lists of retrieved documents. In this section, after a brief illustration of the experimental setting, we report and analyze the results published in the literature. We next discuss alternative evaluation methods.

8.1 Experimental Setting

Most researchers have used, in their experiments, the test collections developed at TREC over the last years. The TREC workshop series is organized in a number of tracks, the ones most relevant to AQE being those that involve searching a static set of documents using new queries (called *topics*); i.e., *ad hoc*, *web*, *robust*, and *terabyte* track. The search tasks evaluated in such tracks mainly differ in the type and size of the collection being searched; the robust track, in addition, explicitly focuses on difficult topics: topics where unexpanded queries achieve poor results. Each collection typically consists of a very large set of documents (drawn from sources such as newswires and the Web), a set of test topics, and manual (human) relevance assessments stating which documents are relevant to which topic. In Table V we report the main document collection statistics.[10]

The most common measure used to evaluate the retrieval effectiveness of the list of documents retrieved in response to a topic is *average precision*. It is defined as the sum of the precision at each relevant document in the list divided by the total number of relevant documents. This measure is computed for each topic and then it is averaged over the set of topics.

8.2 Published Results of AQE Approaches

The data sets summarized in Table V have become a standard benchmark for measuring the retrieval performance of AQE. However, even when referring to the same specific TREC test collection, the published figures are not always directly comparable

[10]Other classical test collections, based on the TREC model, are those provided by CLEF. However, although there are also monolingual search tasks, the emphasis of CLEF is on cross-lingual information retrieval (see Section 4.4).

Table III. A Fine Classification of Several AQE Methods (Continues in Table IV)

Method reference	Data source	Candidate feature extraction method	Candidate feature representation	Feature selection method	Expanded query representation
[Qiu and Frei 1993]	Corpus	All terms in corpus	Single words	Term-concept space	Correlation-based weights
[Voorhees 1994]	WordNet	Query synsets + hyponyms	Single words	Hyponym chain length	Vectors of concept types
[Xu and Croft 1996]	Top-ranked docs + corpus	Adjacent nouns in top-ranked passages	Phrases	Cooccurrence matrix	Rank-based weights
[Mitra et al. 1998]	Top-ranked docs	Independent query terms + adhoc fdbk	Single words	Rocchio weights	Rocchio
[Robertson et al. 1998]	Top-ranked docs	All terms in top-ranked docs	Single words	RSV	Probabilistic reweighting
[Lavrenko and Croft 2001]	Top-ranked docs + corpus	All terms in top-ranked docs	Single words	Relevence model	Interpolated query model
[Zhai and Lafferty 2001a]	Top-ranked docs + corpus	All terms in top-ranked docs	Single words	Mixture model	Interpolated query model
[Carpineto et al. 2001]	Top-ranked docs + corpus	All terms in top-ranked docs	Single words	KLD	Rocchio + KLD scores
[Lam-Adesina and Jones 2001]	Top-ranked docs + corpus	Document summarization	Phrases	RSV	Unweighted terms
[Carmel et al. 2002]	Top-ranked docs + corpus	Lexical affinities	Single words	Splitting rels/nonrels	Rocchio
[Billerbeck et al. 2003]	Query log	Query association	Single words	RSV	Probabilistic reweighting
[Cui et al. 2003]	Query log + corpus	Session-based query-doc correlation	Single words	Probabilistic term-to-term association	Unweighted terms
[Kraft and Zien 2004]	Anchor texts	Adjacent terms in anchor text	Phrases	Median rank aggregation	Unweighted terms
[Liu et al. 2004]	Top-ranked docs + corpus + WordNet	Phrase classification + WordNet concepts	Phrases + single words	Cooccur matrix + disambiguation	Boolean query
[Bai et al. 2005]	Top-ranked docs	Nearby terms in top-ranked docs	Single words	Cooccur matrix + information flow	Interpolated query model

Table IV. A Fine Classification of Several AQE Methods (Continued from Table III)

Method reference	Data source	Candidate feature extraction method	Candidate feature representation	Feature selection method	Expanded query representation
[Graupmann et al. 2005]	Corpus	Web table and form mining	Attribute-value pairs	Association rules	Boolean query
[Collins-Thompson and Callan 2005]	Stemmer + Wordnet + corpus + top-ranked docs	Probabilistic term association network	Single words	Markov chain	Structured query
[Song et al. 2006]	Top-ranked docs + corpus	Keyphrase extraction	Phrases	Information gain + term weighting	DNF of categorized phrases
[Hu et al. 2006]	Corpus	All terms in corpus	Single words	Mutual information	Correlation -based weights
[Riezler et al. 2007]	FAQ training data	Phrases in FAQ answers	Phrases	Statistical machine translation of questions	Unweighted terms
[Bai et al. 2007]	Top-ranked docs + corpus + user domains	Terms + nearby terms	Single words	EM + query classification + mutual information	Query language model combination
[He and Ounis 2007]	Anchor texts + top-ranked docs + corpus	All terms in anchor texts and top-ranked docs	Single words	DFR on fields	Rocchio + combined DFR scores
[Metzler and Croft 2007]	Top-ranked docs + corpus	Markov random fields	Single words + multiword concepts	Maximum likelihood	Expanded query graph
[Lee et al. 2008]	Top-ranked docs + corpus	Clustering of top-ranked docs	Single words	Relevance model	Interpolated query model
[Arguello et al. 2008]	Wikipedia	Anchor texts in top-ranked Wikipedia docs	Phrases	Doc rank + link frequency	Structured query
[Cao et al. 2008]	Top-ranked docs + corpus	All terms in top-ranked docs	Single words	Term classification	Interpolated query model
[Xu et al. 2009]	Wikipedia	All terms in top-ranked articles	Single words	Relevance model	Interpolated query model

Table V. Overview of TREC Collections. The Meanings of the Acronyms are the Following: WSJ = Wall Street Journal, AP = Associated Press newswire, ZIFF = Computer Select Articles (Ziff-Davis), FR = Federal Register, DOE = Abstracts of U.S. Department of Energy publications, SJMN = San Jose Mercury News, PAT = U.S. Patents, FT = Financial Times, CR = Congressional Record, FBIS = Foreign Broadcast Information Service

TREC collection	Description	Size (gigabytes)	Number of docs	Mean number of terms per doc
Disk 1	WSJ (1986-1989), AP (1989) FR (1989), ZIFF, DOE	1.2	510,637	348
Disk 2	WSJ (1990-1992), AP (1988) FR (1989), ZIFF	0.8	231,219	555
Disk 3	SJMN (1991), AP (1990) ZIFF, PAT (1993)	1.1	336,310	481
Disk 4	FT (1991-1994), FR (1994) CR (1993)	1.1	293,710	547
Disk 5	FBIS, the LA Times	0.9	262,367	535
WT10g	1997 crawl of the Internet Archive	10	1,692,096	412
GOV2	2004 crawl of .gov domain	446	25,205,179	691

because the experiments have sometimes been carried out in subtly different conditions. To enable cross-comparison, we considered only the experiments performed

(a) on the full set of documents;
(b) on the full set of topics;
(c) using the title-only description of the topics;
(d) using the mean average precision as evaluation measure.

The results published in the literature have been summarized in Table VI. In addition to the average precision of the single AQE methods, for each test collection we listed the best baseline (a run of the ranking system without AQE) and true relevance feedback when available. In particular, the penultimate row contains the best performance of unexpanded queries (of those reported in the papers associated with the corresponding column), while the last row shows the true relevance feedback performance, taken from Wong et al. [2008] for the TREC6-7-8 collections (making use of Rocchio + Chi-square scores) and from Lee et al. [2008] for the other collections (making use of cluster-based resampling). These latter figures provide upper-bound performance oneach collection (at least for AQE methods based on top retrieved documents), when we are able to choose better pseudo-relevant documents, approaching true relevant documents.

First of all it should be noted that the best absolute results (displayed in bold) varied widely across the various test collections. This phenomenon is evident looking at the results achieved by the same AQE method for the different collections on which it was tested. For instance, the performance of "information flow" ranged from 0.266 (on TREC-1) to 0.394 (on TREC-3). The variability of results can be explained considering that the test collections have very different characteristics; e.g., in terms of size, noise, heterogeneity of contents, difficulty of topics, and so on.

The AQE methods in Table VI belong to several categories described in the preceding. The best results were achieved by four methods, namely "information flow" [Bai et al. 2005] on TREC1-2-3, "query contexts" on TREC7-8 [Bai et al. 2007], "phrases + WordNet" [Liu et al. 2004] on TREC9-10-12, and "Markov random fields" [Metzler and Croft 2007] on TREC13 (Robust track)-14. Interestingly, each was

Table VI. Mean Average Precision of Several AQE Methods on Comparable Test Collections. An * Marks Results that were Significantly Different from the Baseline of No Expansion (Shown in Parenthesis) at the 0.05 Level, Usually According to the Wilcoxon Signed Rank Test. The Best AQE Results for Each Collection are Displayed in Bold. The References are the Following.
(1) [Xu and Croft 2000]; (2) [Zhai and Lafferty 2001a]; (3) [Amati et al. 2001], [Carpineto et al. 2002], [Amati et al. 2004]; (4) [Billerbeck et al. 2003]; (5) [Liu et al. 2004]; (6) [Collins-Thompson and Callan 2005]; (7) [Bai et al. 2005]; (8) [Song et al. 2006]; (9) [Winaver et al. 2007]; (10) [Collins-Thompson and Callan 2007]; (11)[Metzler and Croft 2007]; (12) [Bai et al. 2007]; (13) [He and Ounis 2007]; (14) [Lee et al. 2008]; (15) [Xu et al. 2009]; (16) [Wong et al. 2008]

	trec1 1992	trec2 1993	trec3 1994	trec5 1996	trec6 1997	trec7 1998	trec8 1999	trec9 2000	trec10 2001	trec12 2003	trec13 2004	trec13 2004	trec14 2005	trec15 2006
track	ad hoc	ad hoc	ad hoc	ad hoc	ad hoc	ad hoc	ad hoc	web	web	robust	robust	tera byte	tera byte	tera byte
topics	51 - 100	101- 150	151- 200	251- 300	301- 350	351- 400	401- 450	451- 500	501- 550	301-450 601-650	301-450 601-700	701- 750	751- 800	801- 850
collection	disks 1,2	disks 1...3	disks 1...3	disks 1...4	disks 1...5	disks 1...5	disks 1...5	WT10g (GOV)		disks 1...5	disks 1...5	GOV2		
LCA(1)				**.215** (.210)										
mixture model(2)		.296 (.210)					.282 (.256)							
KLD(3)							.288* (.224)	.187* (.163)	.222 (.178)	.250 (.228)				
association queries(4)								.223* (.189)	.189* (.148)					
phrases + WordNet(5)								**.261** (.187)	**.275** (.183)	**.268** (.189)				
Markov chain(6)	.228 (.214)						.294 (.280)		.227 (.209)					

Table VI. Cont.

information flow[7]	**.266** (.201)	**.318** (.234)	**.394** (.310)											
keyphrases[8]				.198 (.162)	**.211** (.179)	.245 (.222)								
model selection[9]	.238		.303				.270							
model combination[10]	.240 (.181)					.216 (.189)	.226 (.203)		.194 (.174)					
Markov random fields[11]			.269* (.207)						.226* (.186)		**.360*** (.292)		**.392*** (.323)	
query contexts[12]		.248* (.157)				**.246*** (.165)	**.302*** (.238)							
field combination[13]								.199	.242			**.263**	.348	
cluster-based resampling[14]			290* (.207)						.235* (.186)		.351* (.292)		.380* (.325)	
Wikipedia [15]	.162 (.142)							.209* (.183)			.290* (.253)	.339* .296		
best baseline	.214	.234	.310	.210	.179	.222	.280	.189	.209	.232	.292		.325	
true relevance feedback[14],[16]			.425		.555	.534	.488		.403		.535		.431	

consistently better than other methods across all or most tested collections on which it was tested. One thing common to these four methods is that they explicitly took into account term-dependency, although using different techniques; in addition, they primarily made use of top retrieved documents, possibly combined with other sources of evidence, and were built on top of very effective baseline ranking systems.

It is important to note that such findings need to be taken with caution because in Table VI, we listed the absolute overall performance of the system, including, but not limited to, the AQE component. An effective AQE method will clearly yield poor results when combined with an ineffective basic IR system, and vice versa. In fact, the underlying ranking methods employed in the experiments were usually very different and never exactly the same. The baseline performance figures, when reported in the papers, presented considerable variations. The final results achieved in [Bai et al. 2005], for instance, greatly benefitted from a very high baseline performance (e.g., 3107 on TREC-3), even superior to that of the other methods with AQE. We should also consider that even when performing strict single-word indexing and using the same weighting function for document-ranking, there are a number of system-specific factors that can significantly alter the final retrieval performance, including document parsing, stop wording, and stemming. For instance, the removal of spurious, low-frequency words at indexing time from the TREC-9 and TREC-10 collections was observed to be highly beneficial because it reduced the number of typographical errors in documents, which is one of the causes of poor query expansion in noisy collections. There is another issue that can complicate interpretation of results, namely, that the parameters involved in each AQE method might have been optimized using training data or other types of data not always readily or widely available.

8.3 Other Evaluation Methods

To address the shortcomings of the classical AQE evaluation method based on overall change in performance, a few new approaches have recently been proposed. In Custis and Al-Kofahi [2007], the idea is to measure the specific capability of the AQE component in overcoming query-document term mismatch by purposefully degrading the quality of the queries with respect to their relevant documents. In practice, query terms are removed from relevant documents one by one in order of importance (e.g., from highest-to-lowest inverse document frequency) and the performance of the IR systems being evaluated (with or without query expansion) is measured in the standard manner on the altered collections. This approach can be very useful when the use of technical synonyms is the main issue for unsatisfactory information retrieval, as with some domain-specific test collections. Another alternative evaluation strategy, based on the quality of query refinement terms, consists of measuring the degree to which such terms, when used as queries, are capable of retrieving different query aspects or subtopics [Nallapati and Shah 2006]. This latter approach requires labeled documents for the subtopics underlying the query's topic.

Besides evaluating the average retrieval performance, it is also important to consider the robustness of the system. It is well known that the performance of AQE presents a marked variability across queries. In particular, while the majority of queries are improved, some are hurt. Evaluation of robustness has thus become common practice. The standard measure is the *robustness index* (*RI*), defined as the ratio of the difference between the number of queries helped and of those hurt by AQE, to the total number of queries. On TREC data, the fraction of negatively affected queries is of the order of 25% ($RI = 0.5$), if we use the same AQE method across several collections (e.g., Metzler and Croft [2007], Collins-Thompson [2009]).

9. COMPUTATIONAL EFFICIENCY

The total time necessary per performing AQE is the sum of two factors, namely the cost of generating expansion features and the increased cost of evaluating the expanded query against the collection, due to its larger size.

In practice, the latter factor is the most critical one. Consider that most ranking systems have a common architecture based on inverted lists, one for each each term in the collection, where each inverted list specifies which documents that particular term occurs in, usually with a precomputed per-term score. At query time, the system retrieves the inverted list of each query term and updates the score accumulators of the documents present in each list. As query terms are processed one at a time, the execution time of a ranked query is almost linearly dependent on the number of its terms, as also confirmed by experimental observations. For instance, in Billerbeck [2005] and Lavrenko and Allan [2006], AQE runs with sizes of practical interest (ten–twenty words) were found to be much slower than those with original queries, approximately by a factor of ten, yielding final response times on the order of hundreds of milliseconds.

The techniques that have been developed for increasing the efficiency of evaluating ranked queries are based on reducing the number and portion of inverted lists that need to be processed in order to service a query; see e.g. Witten et al. [1999] and Billerbeck [2005]. Documents that are likely to score high are considered with higher priority, and the processing is halted before the whole update has taken place; e.g., as soon as a certain percentage of documents have been given entries in the accumulator table for the current query. The two main strategies for implementing this priority ranking are, (1) evaluating query terms in order of their importance (e.g., by their inverse document frequency), and (2) sorting the documents in the inverted list of a particular term by their relevance to the term (e.g, by their within-document term frequency), followed by parallel execution of ordered inverted lists.

An interesting refinement of such techniques is *top-k query processing* [Theobald et al. 2004, 2005]. The algorithm operates on the same score-sorted index as the previous, but in addition, it makes use of score-distribution statistics to estimate aggregated scores of candidates and to perform early candidate pruning (when the probability of being in the top-k results drops belowan acceptable error threshold). Theoretical and experimental evidence suggests that the use of these approximated faster ranking techniques results in very limited, or even no, degradation of retrieval effectiveness over unapproximated ranking.

Rather than relying on pruning mechanisms, one can try to optimize the full execution of AQE. A recent method [Bast and Weber 2006] suggests precomputing a block-based index, where a block is the union of the inverted lists associated with words that are lexically close, and then adding this information as artificial words to the index for direct use at query time. This index structure explicitly stores expanded inverted lists; it allows faster query processing due to a reduction of random accesses to atomic inverted lists. A similar technique has been applied to advertisement search, treating bid phrases as keyword queries and ads as documents, and placing together in a same block bid, phrases that shared a common prefix [Wang et al. 2009].

Another method for improving the efficiency of unapproximated AQE, although restricted to the language modeling framework, is described in Lavrenko and Allan [2006]; the problem of giant queries is improved by moving some of the computational effort to indexing time via computation of a particular document similarity matrix. Overall, the utilization of the techniques illustrated in this section considerably increased the efficiency of query evaluation, with gains of up to a factor of ten over the traditional inverted index.

Turning to the cost of generating expansion features, it may have a very limited impact on the final response time of the system for several AQE techniques such as extraction of expansion features from query logs and anchor texts, construction of similarity thesauri, and word stemming. The reason is that all possible expansion features are usually generated in advance (e.g., at indexing time, or offline with respect to the underlying ranking system) and the computation left at query time is reduced to selecting those that are appropriate to the current query. The main efficiency concern for such techniques is rather that they may not scale well to large volumes of data due to their inherent complexity (e.g., clustering-based methods grow quadratically with the collection size); this aspect is difficult to evaluate, given the lack of experimental analyses and because the relevant literature is somewhat elusive about the whole efficiency issue.

Of the AQE techniques summarized in the taxonomy in Figure 4, only the query-specific ones raise specific efficiency issues at query time, mainly due to their reliance on a first-pass retrieval. The major bottleneck is fetching the full-text top documents after they have been ranked according to the original query, because these documents are usually stored on disk and disk access times are much slower than memory access times. A more efficient approach is proposed in Billerbeck and Zobel [2004b], making use of short document summaries to be kept in main memory in the form of a set of terms with the highest *tf-idf* values. During querying, all terms in thesummaries that have been ranked against the original query are then used forsourcing expansion terms, thus bypassing disk access altogether and also avoiding the need of parsing the raw documents. Another possibility is to use an external source such as a Web search engine. Downloading the full documents from the Web would clearly be impractical, but using the search result pages is an appealing alternative [Kwok et al. 2004; Yin et al. 2009]. Expansion based on snippetstakes advantage of the engine's large-scale query processing and results-cachinginfrastructure, but it may be subjected to technical limitations (e.g., maximum number of fetched results per query, maximum number of queries per day).

9.1 Which AQE Method is Best?

In general, linguistic techniques are considered less effective than those based on statistical analysis, whether global or local, because they require almost exact word sense disambiguation, but statistical analysis may not always be applied (e.g., when good-expansion terms do not frequently co-occur with the query terms). Of the statistical techniques, local analysis seems to perform better than corpus analysis because the extracted features are query specific, while methods based on Web data (query logs or anchor texts) have not yet been systematically evaluated or compared with the others on standard test collection. The results shown in Table VI confirm this perspective, although they suggest that the single AQE paradigms have a high degree of complementarity that should be exploited to maximize retrieval performance.

From the point of view of computational efficiency, query-specific techniques need a double run at query time while other forms of AQE are mostly performed in an offline stage, but the inherent complexity of the latter techniques may prevent their application in high dimensionality domains. Besides effectiveness and efficiency, there are other points that should be considered. Query-specific techniques are dependent on the quality of the first-pass retrieval, corpus-specific techniques are not suitable for dynamic document collections, linguistic techniques and methods based on analysis of query logs or hyperlinks make use of data that are not always available or suitable for the IR task at hand. Finally, some AQE techniques require the capability of evaluating structured expanded queries.

To summarize, there is a wide range of AQE techniques that present different features and are mostly useful or applicable in certain situations. The best choice depends on the evaluation of a number of factors, including type of collection being searched, availability and characteristics of external data, facilities offered by the underlying ranking system, type of queries, and efficiency requirements.

10. CRITICAL ISSUES

In this section we discuss three key issues that pose obstacles for a widespread adoption of AQE in a wider range of operational search systems: parameter setting, efficiency, and usability.

10.1 Parameter Setting

All AQE techniques rely on several parameters. For instance, for a typical pseudo-relevance feedback method it is necessary to choose the number of pseudo-relevant documents, the number of expansion terms, and the λ balance coefficient for query reformulation. The retrieval performance of the overall method is usually markedly dependent on the parameter setting.

The standard approach is to use fixed values for key parameters, determinedby fine tuning on test collections. But there are two main drawbacks. The first is that a fixed value for all queries is probably not the best choice. Queries have different characteristics in terms of length, difficulty, verbosity, ambiguity, and goal, and they should receive an individual treatment. Several experiments (e.g., Carpineto et al. [2002], Billerbeck [2005]) showed that the use of fixed parameter values results in a heavy penalization of average retrieval performance, compared to that theoretically obtainable with optimal query-based AQE, and it is probably one of the main reasons for the unsatisfactory robustness of AQE. The second problem is that while such a tuning is standard in benchmarkslike TREC, it becomes very difficult for real search applications with highly dynamic corpora, unpredictable queries,and continuously evolving search tasks (e.g., Web IR, intranets,digital libraries, Web communities, etc.)

This calls for automatic and self-adaptive query-based parameter setting. We will see in the next section that this issue has started to be investigated but there are still many challenges.

10.2 Efficiency

The efficient evaluation of queries is essential for IR systems such as web search enginesthat need to deliver real-time results to a very large numbers of users. While the expansion feature generation stage can be carried out efficiently, the successive execution of the expanded query may become too slow, as discussed in Section 9. This slowdown may prevent theadoption of AQE for real retrieval systems. It is also harmful to research because far fewer runs fit into a particular window of time [Lavrenko and Allan 2006]. Faster AQE techniques would allow researchers to carry outmore experiments and interactive studies to better understand the applicability and limitations of this methodology.

There are three possible ways to address this issue:

— limit expansion features to a few important items and then rank the expanded query in a standard way,
— allow for a possibly large number of expansion features, but prune features and documents that are unlikely to lead to an improved result when ranking the expanded query,
— use efficient index structure (for applications when it is possible) that support nearly full document ranking against nearly full expanded queries.

The adoption of such approximated techniques usually involves a moderate trade-off between speed and retrieval performance, although their overall adequacy ultimately depends on the requirements posed by the search application.

10.3 Usability

Usability is probably another critical issue, although it has not received much attention so far. AQE acts like a black box employing hidden features that may considerably complicate the interpretation of the logic used by the system to deliver results. For instance, some Web users may be unsatisfied finding documents (even relevant ones) that do not contain the terms in their query. This happens sometimes, using AQE. For example, a document may be returned because the anchor texts pointing to it contain the query terms, or because a query term is subsumed by a more general term in the document, according to a given ontology. When users obtain a result set that they find inadequate, they have no explanation for why certain results not containing the original query terms were ranked high so they have no easy way to improve the query.

A simple method to reduce the lack of transparency and increase user control over the relationships between query and results is to display the list of expansion features used by the system for ranking the documents. A more comprehensive approach would be not only to show why certain search results have been obtained, but also to allow some form of manipulation of query and results on the part of the user; e.g., revising the expanded query, zooming in on results that are related to some expansion features, and so on.

It is known [Ruthven 2003] that expert users are capable of taking full advantage of a query refinement feature, whereas pure AQE is better for non-expert users. Hybrid strategies that integrate AQE and interactive search facilities might be more effective for all types of users, but they have not been much investigated so far. A notable exception is Bast et al. [2007], where the individual expansion terms and the number of their associated hits are displayed automatically after each keystroke in the search box, together with the best hits. A similar search paradigm has recently been followed for improving content-based visual retrieval, using pairs formed by a refinement keyword and its associated representative images as single expansion features [Zha et al. 2009]. Overall, the usability issue in AQE needs more research.

11. RESEARCH DIRECTIONS

Most current research effort aims at improving the retrieval effectiveness and robustness of AQE. In this section we focus on three relatively well-established topics: selective AQE, evidence combination, and active feedback. Other directions that are being investigated are attempts to integrate personal [Chirita et al. 2007] and negative relevance feedback [Bernardini and Carpineto 2008; Wang et al. 2008] information in the AQE framework, as well as more sophisticated forms of implicit user feedback such as eye tracking [Buscher et al. 2008].

11.1 Selective AQE

Selective AQE aims to improve query expansion with decision mechanisms based on the characteristics of queries. Based on the observation that some queries are hurt by expansion, one simple strategy is to disable AQE if the query can be predicted to perform poorly. However, it is not obvious which properties make a query suitable/unsuitable for AQE. For instance, easy (difficult) queries do not necessarily produce better (worse) performance, as there is no clear correlation between the average precision that the original query achieves and by how much AQE improves average precision [Billerbeck and Zobel 2003; Carpineto et al. 2001; He and Ounis 2009b].

The best known predictive function, termed *clarity score* [Cronen-Townsend and Croft 2002], is the Kullback-Leibler divergence between the query model, estimated from the top-ranked documents, and the collection model. In principle, the higher the divergence, the better the retrieval performance that AQE can provide. Experimentally, however, a straightforward use of the clarity score is not always beneficial. Perhaps, a more effective strategy is to use the difference between the clarity score of theinitial query and that of the expanded query, yielding performance results comparable to ranking without AQE on the worst queries (those hurt by expansion) and better than conventional AQE on the whole set of queries [Amati et al. 2004].

Rather than just disabling AQE when its application is deemed harmful, it may be more convenient to apply different expansion strategies according to the type of query. An interesting form of query-dependent AQE is presented in Xu et al. [2009] using Wikipedia pages. Queries are classified in three types: (1) entity queries, if they match the title of an entity or redirect page, (2) ambiguous queries, if they match the title of a disambiguation page, and (3) broader queries in all other cases. For each type of query, a different method of AQE is then carried out. Another approach that exploits a similar idea is described in Fujii [2008]. Queries are classified as either navigational or informational, making use of anchor-link distribution. Navigational queries are then handled by a particular anchor-based retrieval model that expands anchor terms with their synonyms. Focused expansions have also been applied in a federated search setting, producing specific queries for each source [Shokouhi et al. 2009].

11.2 Evidence Combination

Distinct AQE methods usually produce different refinements, with low to moderate overlap [Kraft and Zien 2004]. Even when the overlap is large, the ordered sets of expansion features may be largely uncorrelated [Carpineto et al. 2002]. If the refinements suggested by the single methods are, on the whole, equally useful (e.g., they result in comparable average performance over a set of queries), one can try to combine the most effective refinements at the individual query level. This strategy often works fairly well, with the combined method improving over all single methods.

Several combination methods have been proposed. Two approaches, already mentioned, consist of selecting the most common terms of those produced by multiple term-ranking functions [Carpineto et al. 2002], or classifying as relevant or non-relevant the terms produced by the same term-ranking function with different document samples [Collins-Thompson and Callan 2007]. In He and Ounis [2007], the focus is on improving the quality of query term reweighting, rather than choosing the best terms, by taking a linear combination of the term frequencies in three document fields (title, anchor texts, body). All these combination methods were applied as an improvement of pseudo-relevance feedback.

Linguistically-oriented AQE techniques can also greatly benefit from a combined approach due to data sparsity: general-purpose resources are limited in coverage and depth, but they can complement co-occurrence relations when the latter evidence is not available or reliable. In Liu et al. [2004], WordNet concepts (synonyms and hyponyms) are combined by heuristic rules with other expansion features extracted using global and local statistical methods. In Bai et al. [2007], multiple query expansion models are employed (using an external ontology, the whole collection, and the top retrieved documents) and then they are combined by interpolation to yield the final expanded query model

$$p(t|\theta_q) = \sum_i \alpha_i \; p(t|\theta_q^i), \tag{14}$$

with $\sum_i \alpha_i = 1$. The best settings of such mixture weights confirmed that all models affected the final performance, but the pseudo-relevance feedback model was largely the most effective one. Another two approaches that combine complex linguistic and statistical features are discussed in Collins-Thompson and Callan [2005] and Metzler and Croft [2007]; both were reviewed in the preceding sections.

11.3 Active Feedback

In query-specific AQE techniques, treating the top documents as relevant is often not the best strategy. For example, if the top documents have very similar contents, their cumulative benefit will not be very different from that attainable from any one of them. The main approach for choosing more informative feedback documents is to emphasize their diversity. Several techniques have been proposed, such as reranking documents based on independent query concepts [Mitra et al. 1998], using cluster centroids or ranking gaps [Shen and Zhai 2005], skipping redundant documents [Sakai et al. 2005], estimating uncertainty associated with a feedback model [Collins-Thompson and Callan 2007], and choosing documents that appear in multiple overlapping clusters [Lee et al. 2008]. Diversity is always combined, explicitly or implicitly, with relevance. In Xu and Akella [2007], a more comprehensive framework is presented, which integrates relevance, diversity, and density, where density is measured as the average distance of a document from all other documents.

In He and Ounis [2009a], the authors present a machine learning approach to active feedback, analogous to that used in Cao et al. [2008] to select relevant expansion terms. They classify the top-retrieved documents as good or bad, using various features such as the distribution of query terms in the document and the proximity between the expansion terms and the original query terms in the document. To train the classifier, they use top-retrieved documents labeled as good or bad depending on whether they improve or hurt retrieval performance when used as feedback documents.

Usually, it is assumed that the collection from which to extract the documents for AQE is fixed. Selection from a multidatabase corpus is an interesting larger-scale form of document selection. It turns out that analyzing the databases separately can be better than treating the corpus as one large database, with substantial improvements if the best database is chosen [Gauch et al. 1999]. The most appropriate database can be chosen by running the query against the individual databases and analyzing the search results [Gauch et al. 1999] or by more efficient, preretrieval query performance predictors [Hauff et al. 2008; He and Ounis 2007]. Besides optimizing the choice of the best feedback documents, one can also focus on their best parts. An earlier textual form of this approach is passage selection [Xu and Croft 1996]; other AQE methods suitable for Web pages involve more sophisticated features such as visual clues [Yu et al. 2003] or tables and forms [Graupmann et al. 2005].

12. CONCLUSIONS

Although there is no silver bullet for the vocabulary problem in IR, AQE has the potential to overcome one of the main limitations of current search systems usage: the reluctance and the difficulty of users in providing a more precise description of their information needs. In the last ten years, AQE has made a big leap forward, by leveraging diverse data sources and inventing more principled and effective methods. Nowadays a spectrum of techniques is available (e.g., linguistic, corpus-specific, query-specific, based on search logs, and on Web data) that cater to different requirements in terms of query type, computational efficiency, availability of external data, and characteristics of the underlying ranking system.

The advance of AQE techniques has been confirmed by a number of experimental tests on classical benchmarks. Remarkable improvements in average retrieval effectiveness have been reported in most evaluation studies, with gains not only in recall but also in precision, at least for some types of queries.

In spite of such good results, AQE still suffers from drawbacks that have limited its deployment as a standard component in search systems. The key aspects that need to be improved are the robustness of retrieval performance, the automatic setting of parameters, the computational efficiency of executing larger queries, and the usability of an IR system implementing AQE. These limitations have started to be addressed in recent research, together with the exploration of new directions.

Among the most promising trends are the development of AQE methods that explicitly take into account term dependency (e.g., through a combination of statistical and linguistic techniques), the utilization of search query languages that allow for structured expanded queries, and the injection of interactive facilities into the basic AQE framework. Hybrid methods achieved the best results on the experimental benchmarks and seem, in principle, more robust with respect to variation of queries, document collections, and users. Equally important are the exploration and learning of adaptive techniques. Query-dependent criteria can be used to optimize the amount of expansion, the type of reformulation, the setting of parameters, and the selective application of AQE.

In summary, AQE may be at a turning point after about forty years of research. It has reached a level of scientific maturity and there are signs that it is moving beyond its experimental status and being adopted in operating systems. This article will hopefully help to make AQE better known and more widely accepted and used in the search market.

ACKNOWLEDGMENTS

We are very grateful to three anonymous reviewers for their excellent comments and suggestions.

REFERENCES

AGICHTEIN, E., LAWRENCE, S., AND GRAVANO, L. 2004. Learning to find answers to questions on the Web. *ACM Trans. on Internet Technol. 4*, 2, 1299–162.

AGIRRE, E., ANSA, O., ARREGI, X., DE LACALLE, M. L., OTEGI, A., SARALEGI, X., AND SARAGOZA, H. 2009. Elhuyar-ixa: Semantic relatedness and cross-lingual passage retrieval. In *Proceedings of CLEF*. Springer.

AGIRRE, E., DI NUNZIO, G. M., MANDL, T., AND OTEGI, A. 2009. Clef 2009 ad hoc track overview: Robust–wsd task. In *Proceedings of CLEF*. Springer.

AGRAWAL, R., IMIELINSKI, T., AND SWAMI, A. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM Press, 207–216.

ALLAN, J. 1996. Incremental relevance feedback for information filtering. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 270–278.

AMATI, G. 2003. Probabilistic models for information retrieval based on divergence from randomness. Ph.D. thesis, Department of Computing Science, University of Glasgow, UK.

AMATI, G., CARPINETO, C., AND ROMANO, G. 2001. FUB at TREC-10 Web Track: A probabilistic framework for topic relevance term weighting. In *Proceedings of the 10th Text REtrieval Conference (TREC'10)*. NIST Special Publication 500–250. National Institute of Standards and Technology (NIST), Gaithersburg, MD, 182–191.

AMATI, G., CARPINETO, C., AND ROMANO, G. 2003. Comparing weighting models for monolingual information retrieval. In *Proceedings of the 4th Workshop of the Cross-Language Evaluation Forum (CLEF'03)*. Springer, 310–318.

AMATI, G., CARPINETO, C., AND ROMANO, G. 2004. Query difficulty, robustness, and selective application of query expansion. In *Proceedings of the 26th European Conference on Information Retrieval (ECIR'04)*. Springer, 127–137.

ANDERSON, J. R. 1983. A spreading activation theory of memory. *J. Verbal Learn. Verbal Behav. 22*, 261–295.

ARGUELLO, J., ELSAS, J. L., CALLAN, J., AND CARBONELL, J. G. 2008. Document representation and query expansion models for blog recommendation. In *Proceedings of the 2nd International Conference on Weblogs and Social Media*. AAAI Press, 10–18.

ATTAR, R. AND FRAENKEL, A. S. 1977. Local feedback in full-text retrieval systems. *J. ACM 24*, 3, 397–417.

BAEZA-YATES, R. AND RIBEIRO-NETO, B. 1999. *Modern Information Retrieval*. Addison Wesley.

BAI, J., NIE, J.-Y., AND CAO, G. 2006. Context-dependent term relations for information retrieval. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 551–559.

BAI, J., SONG, D., BRUZA, P., NIE, J.-Y., AND CAO, G. 2005. Query expansion using term relationships in language models for information retrieval. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. ACM Press, 688–695.

BAI, J., NIE, J.-Y., CAO, G., AND BOUCHARD, H. 2007. Using query contexts in information retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 15–22.

BALLESTEROS, L. AND CROFT, W. B. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 84–91.

BALLESTEROS, L. AND CROFT, W. B. 1998. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 64–71.

BAST, H. AND WEBER, I. 2006. Type less, find more: fast autocompletion search with a succinct index. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 364–371.

BAST, H., MAJUMDAR, D., AND WEBER, I. 2007. Efficient interactive query expansion with complete search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 857–860.

BEEFERMAN, D. AND BERGER, A. 2000. Agglomerative clustering of a search engine query log. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, 407–416.

BELKIN, N. J. AND CROFT, W. B. 1992. Information filtering and information retrieval: Two sides of the same coin? *Comm. ACM 35*, 12, 29–38.

BERNARDINI, A. AND CARPINETO, C. 2008. Fub at trec 2008 relevance feedback track: extending rocchio with distributional term analysis. In *Proceedings of TREC-2008*. National Institute of Standards and Technology, Gaithersburg, MD, USA.

BERNARDINI, A., CARPINETO, C., AND D'AMICO, M. 2009. Full-subtopic retrieval with keyphrase-based search results clustering. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, 206–213.

BHOGAL, J., MACFARLANE, A., AND SMITH, P. 2007. A review of ontology based query expansion. *Info. Process. Manage. 43*, 4, 866–886.

BILLERBECK, B. 2005. Efficient query expansion. Ph.D. thesis, RMIT University, Melbourne, Australia.

BILLERBECK, B. AND ZOBEL, J. 2003. When query expansion fails. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. ACM Press, 387–388.

BILLERBECK, B. AND ZOBEL, J. 2004a. Questioning query expansion: An examination of behaviour and parameters. In *Proceedings of the 15th Australasian Database Conference.* Vol. 27, Australian Computer Society, 69–76.

BILLERBECK, B. AND ZOBEL, J. 2004b. Techniques for efficient query expansion. In *Proceedings of the String Processing and Information Retrieval Symposium*. Springer, 30–42.

BILLERBECK, B. AND ZOBEL, J. 2005. Document expansion versus query expansion for ad-hoc retrieval. In *Proceedings of the 10th Australasian Document Computing Symposium*. Australian Computer Society, Sydney, Australia, 34–41.

BILLERBECK, B., SCHOLER, F., WILLIAMS, H. E., AND ZOBEL, J. 2003. Query expansion using associated queries. In *Proceedings of the 12th ACM International Conference on Information and Knowledge Management*. ACM Press, 2–9.

BILOTTI, M., KATZ, B., AND LIN, J. 2004. What works better for question answering: Stemming or morphological query expansion? In *Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR'04*.

BODOFF, D. AND KAMBIL, A. 1998. Partial coordination. I. The best of pre-coordination and post-coordination. *J. Amer. Soc. Info. Sciences 49*, 14, 1254–1269.

BRODER, A. 2002. A taxonomy of web search. *ACM SIGIR Forum 36*, 2, 3–10.

BRODER, A., CICCOLO, P., E.GABRILOVICH, JOSIFOVSKI, V., METZLER, D., RIEDEL, L., AND YUAN, J. 2009. Online expansion of rare queries for sponsored search. In *Proceedings of the 18th international conference on World Wide Web*. ACM, 511–520.

BUCKLEY, C. AND HARMAN, D. K. 2003. Reliable information access final workshop report. In *Proceedings of the Reliable Information Access Workshop (RIA)*. NRRC, 1–30.

BUCKLEY, C., SALTON, G., ALLAN, G., AND SINGHAL, A. 1995. Automatic query expansion using smart: Trec3. In *Proceedings of the 3rd Text REtrieval Conference (TREC-3)*. NIST Special Publication 500–226. National Institute of Standards and Technology (NIST), Gaithersburg, MD, 69–80.

BUSCHER, G., DENGEL, A., AND VAN ELST, L. 2008. Query expansion using gaze-based feedback on the subdocument level. In *Proceedings of the 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 387–394.

CAO, G., GAO, J., NIE, J.-Y., AND BAI, J. 2007. Extending query translation to cross-language query expansion with markov chain models. In *Proceedings of the 16th Conference on Information and Knowledge Management (CIKM'07)*. ACM Press.

CAO, G., GAO, J., NIE, J.-Y., AND ROBERTSON, S. 2008. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 243–250.

CARMEL, D., FARCHI, E., PETRUSCHKA, Y., AND SOFFER, A. 2002. Automatic query refinement using lexical affinities with maximal information gain. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 283–290.

CARPINETO, C. AND ROMANO, G. 2004. *Concept Data Analysis: Theory and Applications*. John Wiley & Sons.

CARPINETO, C., DE MORI, R., ROMANO, G., AND BIGI, B. 2001. An information theoretic approach to automatic query expansion. *ACM Trans. Info. Syst. 19*, 1, 1–27.

CARPINETO, C., ROMANO, G., AND GIANNINI, V. 2002. Improving retrieval feedback with multiple term-ranking function combination. *ACM Trans. Info. Syst. 20*, 3, 259–290.

CARPINETO, C., OSIŃSKI, S., ROMANO, G., AND WEISS, D. 2009. A survey of Web clustering engines. *ACM Comput. Surv. 41*, 3.

CHANG, Y., OUNIS, I., AND KIM, M. 2006. Query reformulation using automatically generated query concepts from a document space. *Info. Process. Manage. 42*, 2, 453–468.

CHEN, L., L'ABBATE, M., THIEL, U., AND NEUHOLD, E. J. 2004. Increasing the customers choice: Query expansion based on the layer-seeds method and its application in e-commerce. In *Proceedings of the IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'04)*. IEEE Computer Society, 317–324.

CHIRITA, P.-A., FIRAN, C. S., AND NEJDL, W. 2007. Personalized query expansion for the web. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 7–14.

CHU, W. W., LIU, Z., AND MAO, W. 2002. Textual document indexing and retrieval via knowledge sources and data mining. *Comm. Institute of Info. Comput. Machinery 5*, 2.

CHURCH, K. AND HANKS, P. 1990. Word association norms, mutual information and lexicography. *Computat. Linguist. 16*, 1, 22–29.

CHURCH, K. AND SMYTH, B. 2007. Mobile content enrichment. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*. ACM Press, 112–121.

COLLINS-THOMPSON, K. 2009. Reducing the risk of query expansion via robust constrained optimization. In *Proceedings of the 18th Conference on Information and Knowledge Management (CIKM'09)*. ACM Press, 837–846.

COLLINS-THOMPSON, K. AND CALLAN, J. 2005. Query expansion using random walk models. In *Proceedings of the 14th Conference on Information and Knowledge Management (CIKM'05)*. ACM Press, 704–711.

COLLINS-THOMPSON, K. AND CALLAN, J. 2007. Estimation and use of uncertainty in pseudo-relevance feedback. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 303–310.

CRABTREE, D., ANDREAE, P., AND GAO, X. 2007. Exploiting underrepresented query aspects for automatic query expansion. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, 191–200.

CRESTANI, F. 1997. Application of spreading activation techniques in information retrieval. *Artif. Intell. 11*, 6, 453–482.

CRONEN-TOWNSEND, S. AND CROFT, W. B. 2002. Quantifying query ambiguity. In *Proceedings of the 2nd International Conference on Human Language Technology Research*. ACM Press, 104–109.

CROUCH, C. AND YANG, B. 1992. Experiments in automatic statistical thesaurus construction. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 77–88.

CUI, H., WEN, J.-R., NIE, J.-Y., AND MA, W.-Y. 2003. Query expansion by mining user logs. *IEEE Trans. Knowl. Data Engin. 15*, 4, 829–839.

CUSTIS, T. AND AL-KOFAHI, K. 2007. A new approach for evaluating query expansion: Query-document term mismatch. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 575–582.

DEERWESTER, S., DUMAIS, S. T., FURNAS, W., LANDAUER, T. K., AND HARSHMAN, R. 1990. Indexing by latent semantic analysis. *J. Amer. Soc. Info. Science 41*, 6, 391–407.

DEMPSTER, A., LAIRD, N., AND RUBIN, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc.* Series B (Methodological) *39*, 1, 1–38.

DIAZ, F. AND METZLER, D. 2006. Improving the estimation of relevance models using large external corpora. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 154–161.

DOSZKOCS, T. E. 1978. AID, an Associative Interactive Dictionary for Online Searching. *Online Rev. 2*, 2, 163–174.

EFRON, M. 2008. Query Expansion and Dimensionality Reduction: Notions of Optimality in Rocchio Relevance Feedback and Latent Semantic Indexing. *Info. Process. Manage. 44*, 1, 163–180.

EFTHIMIADIS, E. N. 1993. A user-centred evaluation of ranking algorithms for interactive query expansion. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 146–159.

EFTHIMIADIS, E. N. 1996. Query expansion. In *Annual Review of Information Systems and Technology*, M. E. Williams Ed., ASIS&T, 121–187.

FINKELSTEIN, L., GABRILOVICH, E., MATIAS, Y., RIVLIN, E., SOLAN, Z., WOLFMAN, G., AND RUPPIN, E. 2002. Placing search in context: The concept revisited. *ACM Trans. Info. Syst. 20*, 1, 116–131.

FITZPATRICK, L. AND DENT, M. 1997. Automatic feedback using past queries: Social searching? In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 306–313.

FLEMMINGS, R., BARROS, J., GERALDO, A. P., AND MOREIRA, V. P. 2009. Bbk-ufrgs@clef2009: Query expansion of geographic place names. In *Proceedings of CLEF*.

FUJII, A. 2008. Modeling anchor text and classifying queries to enhance web document retrieval. In *Proceeding of the 17th International Conference on World Wide Web*. ACM Press, 337–346.

FURNAS, G. W., LANDAUER, T. K., GOMEZ, L. M., AND DUMAIS, S. T. 1987. The vocabulary problem in human-system communication. *Comm. ACM 30*, 11, 964–971.

GAUCH, S., WANG, J., AND RACHAKONDA, S. M. 1999. A corpus analysis approach for automatic query expansion and its extension to multiple databases. *ACM Trans. Info. Syst. 17*, 3, 250–269.

GONG, Z., CHEANG, C.-W., AND U, L. 2006. Multi-term web query expansion using wordnet. In *Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA'06)*. Springer, 379–388.

GONZALO, J., VERDEJO, F., CHUGUR, I., AND CIGARRÄN, J. M. 1998. Indexing with wordnet synsets can improve text retrieval. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*. Association for Computational Linguistics, 647–678.

GRAUPMANN, J., CAI, J., AND SCHENKEL, R. 2005. Automatic query refinement using mined semantic relations. In *Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration (WIRI)*. IEEE Computer Society, 205–213.

HANANI, U., SHAPIRA, B., AND SHOVAL, P. 2004. Information filtering: Overview of issues, research and systems. *User Model. User-Adapt. Interact. 11*, 3, 203–259.

HARABAGIU, S. AND LACATUSU, F. 2004. Strategies for advanced question answering. In *Proceedings of the HLT- NAACL'04 Workshop on Pragmatics of Question Answering*. 1–9.

HARABAGIU, S., MOLDOVAN, D., PASCA, M., MIHALCEA, R., SURDEANU, M., BUNESCU, R., GRJU, R., RUS, V., AND MORARESCU, P. 2001. The role of lexico-semantic feedback in open-domain textual question-answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-01)*. Association for Computational Linguistics, 282–289.

HARMAN, D. K. 1992. Relevance feedback and other query modification techniques. In *Information Retrieval – Data Structures and Algorithms*, W. B. Frakes and R. Baeza-Yates Eds., Prentice Hall, Englewood Cliffs, N. J., 241–263.

HARPER, G. W. AND VAN RIJSBERGEN, C. J. 1978. An evaluation of feedback in document retrieval using co-occurrence data. *J. Documentation 34*, 3, 189–216.

HAUFF, C., HIEMSTRA, D., AND DE JONG, F. 2008. A survey of pre-retrieval query performance predictors. In *Proceedings of the 17th Conference on Information and Knowledge Management (CIKM'08)*. ACM Press, 1419–1420.

HE, B. AND OUNIS, I. 2007. Combining fields for query expansion and adaptive query expansion. *Info. Process. Manage. 43*, 1294–1307.

HE, B. AND OUNIS, I. 2009a. Finding good feedback documents. In *Proceedings of the 18th Conference on Information and Knowledge Management (CIKM'09)*. ACM Press, 2011–2014.

HE, B. AND OUNIS, I. 2009b. Studying query expansion effectiveness. In *Proceedings of the 31th European Conference on Information Retrieval (ECIR'09)*. Springer, 611–619.

HIDALGO, J. M. G., DE BUENAGA RODRÍGUEZ, M., AND PÉREZ, J. C. C. 2005. The role of word sense disambiguation in automated text categorization. In *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems*. Springer, 298–309.

HU, J., DENG, W., AND GUO, J. 2006. Improving retrieval performance by global analysis. In *Proceedings of the 18th International Conference on Pattern Recognition*. IEEE Computer Society, 703–706.

HUANG, C.-C., CHIEN, L.-F., AND OYANG, Y.-J. 2003. Relevant term suggestion in interactive web search based on contextual information in query session logs. *J. Amer. Soc. Info. Science Technol. 54*, 7, 638–649.

HUANG, C.-C., LIN, K.-M., AND CHIEN, L.-F. 2005. Automatic training corpora acquisition through web mining. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, 193–199.

HULL, D. A. 1996. Stemming algorithms: a case study for detailed evaluation. *J. Amer. Soc. Info. Science 47*, 1, 70–84.

IDE, E. 1971. New experiments in relevance feedback. In *The SMART Retrieval System*, G. Salton Ed., Prentice Hall, Englewood Cliffs, N. J., 337–354.

JELINEK, F. AND MERCER, R. L. 1980. Interpolated estimation of markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*. North-Holland, Amsterdam, The Netherlands, 381–397.

JOACHIMS, T., GRANKA, L., PAN, B., HEMBROOKE, H., RADLINSKI, F., AND GAY, G. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Info. Syst. 25*, 2, 7.

JONES, R., REY, B., MADANI, O., AND GREINER, W. 2006. Generating query substitutions. In *Proceedings of the 15th International Conference on World Wide Web*. ACM Press, 387–396.

JONES, S. 1993. A thesaurus data model for an intelligent retrieval system. *J. Info. Science 19*, 3, 167–178.

JONES, S. 1995. Interactive thesaurus navigation: Intelligence rules ok? *J. Amer. Soc. for Info. Science 46*, 1, 52–59.

KAMVAR, M. AND BALUJA, S. 2007. The role of context in query input: Using contextual signals to complete queries on mobile devices. In *Proceedings of the 9th International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM Press, 405–412.

KANAAN, G., AL-SHALABI, R., GHWANMEH, S., AND BANI-ISMAIL, B. 2008. Interactive and automatic query expansion: A comparative study with an application on Arabic. *Amer. J. Appl. Sciences 5*, 11, 1433–1436.

KEKÄLÄINEN, J. AND JÄRVELIN, K. 1998. The impact of query structure and query expansion on retrieval performance. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 130–137.

KHERFI, M. L., ZIOU, D., AND BERNARDI., A. 2004. Image retrieval from the World Wide Web: Issues, techniques, and systems. *ACM Comput. Surv. 36*, 1, 35–67.

KOEHN, P. 2010. *Statistical Machine Translation*. Cambridge University Press.

KRAAIJ, W., NIE, J., AND SIMARD, M. 2003. Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval. *Computat. Linguist. 29*, 3, 381–420.

KRAFT, R. AND ZIEN, J. 2004. Mining anchor text for query refinement. In *Proceedings of the 13th International Conference on World Wide Web*. ACM Press, 666–674.

KROVETZ, R. 1993. Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 191–202.

KROVETZ, R. AND CROFT, W. B. 1992. Lexical ambiguity and information retrieval. *ACM Trans. Info. Syst. 10*, 2, 115–141.

KURLAND, O., LEE, L., AND DOMSHLAK, C. 2005. Better than the real thing?: Iterative pseudo-query processing using cluster-based language models. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 19–26.

KWOK, K. L., GRUNFELD, L., SUN, K. L., AND DENG, P. 2004. TREC2004 robust track experiments using PIRCS. In *Proceedings of the 13th Text REtrieval Conference (TREC-8)*. National Institute of Standards and Technology, Gaithersburg, MD.

LAM-ADESINA, A. M. AND JONES, G. J. F. 2001. Applying summarization techniques for term selection in relevance feedback. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–9.

LATIRI, C. C., YAHIA, S. B., CHEVALLET, J. P., AND JAOUA, A. 2004. Query expansion using fuzzy association rules between terms. In *Proceedings of the 4th International Conference Journées de l'Informatique Messine (JIM'03)*.

LAU, R. Y. K., BRUZA, P. D., AND SONG, D. 2004. Belief revision for adaptive information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 130–137.

LAU, T. AND HORVITZ, E. 1999. Patterns of search: Analyzing and modeling web query refinement. In *Proceedings of the 7th International Conference on User Modeling*. Springer, 119–128.

LAVELLI, A., SEBASTIANI, F., AND ZANOLI, R. 2004. Distributional term representations: an experimental comparison. In *Proceedings of the 16th Conference on Information and Knowledge Management (CIKM'04)*. ACM Press, 615–624.

LAVRENKO, V. AND ALLAN, J. 2006. Realtime query expansion in relevance models. IR 473, University of Massachusetts.

LAVRENKO, V. AND CROFT, W. B. 2001. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 120–127.

LEE, K. S., CROFT, W. B., AND ALLAN, J. 2008. A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of the 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 235–242.

LESK, M. E. 1969. Word-Word Associations in Document Retrieval Systems. *Amer. Documentation 20*, 1, 8–36.

LESK, M. E. 1988. They said true things, but called them by wrong names – vocabulary problems over time in retrieval. In *Proceedings of the Waterloo OED Conference*. ACM Press, 1–10.

LIN, J. AND MURRAY, G. C. 2005. Assessing the term independence assumption in blind relevance feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 635–636.

LIU, S., LIU, F., YU, C., AND MENG, W. 2004. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 266–272.

LIU, Y., LI, C., ZHANG, P., AND XIONG, Z. 2008. A query expansion algorithm based on phrases semantic similarity. In *Proceedings of the International Symposiums on Information Processing*. IEEE Computer Society, 31–35.

LV, Y. AND ZHAI, C. 2009. Adaptive relevance feedback in information retrieval. In *Proceedings of the 18th Conference on Information and Knowledge Management (CIKM'09)*. ACM Press, 255–264.

MACDONALD, C. AND OUNIS, I. 2007. Expertise drift and query expansion in expert search. In *Proceedings of the 16th Conference on Information and Knowledge Management (CIKM'07)*. ACM Press.

MANDALA, R., TAKENOBU, T., AND HOZUMI, T. 1998. The use of wordnet in information retrieval. In *Proceedings of the ACL Workshop on the Usage of WordNet in Information Retrieval*. Association for Computational Linguistics, 31–37.

MANDALA, R., TOKUNAGA, T., AND TANAKA, H. 1999. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 191–197.

MANNING, C. D., RAGHAVAN, P., AND SCH'UTZE, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

MARON, M. E. AND KUHNS, J. L. 1960. On relevance, probabilistic indexing and information retrieval. *J. ACM 7*, 3, 216–244.

MCNAMEE, P. AND MAYFIELD, J. 2002. Comparing cross-language query expansion techniques by degrading translation resources. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 159–166.

MELUCCI, M. 2008. A Basis for Information Retrieval in Context. *ACM Trans. Info. Syst. 26*, 3, Article No 14.

METZLER, D. AND CROFT, W. B. 2007. Latent concept expansion using Markov random fields. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 311–318.

MILLER, G. A., BECKWITH, R. T., FELLBAUM, C. D., GROSS, D., AND MILLER, K. 1990. WordNet: An online lexical database. *Int. J. Lexicography 3*, 4, 235–244.

MILNE, D. N., WITTEN, I. H., AND NICHOLS, D. M. 2007. A knowledge-based search engine powered by wikipedia. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*. ACM Press, 445–454.

MINKER, J., WILSON, G. A., AND ZIMMERMAN, B. H. 1972. An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Info. Stor. Retrieval 8*, 6, 329–348.

MITRA, M., SINGHAL, A., AND BUCKLEY, C. 1998. Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 206–214.

MONTAGUE, M. AND ASLAM, J. 2001. Relevance score normalization for metasearch. In *Proceedings of the 10th International Conference on Information and Knowledge Management*. ACM Press, 427–433.

NALLAPATI, R. AND SHAH, C. 2006. Evaluating the quality of query refinement suggestions in information retrieval. IR 521, University of Massachusetts.

NATSEV, A., HAUBOLD, A., TEŠIĆ, J., XIE, L., AND YAN, R. 2007. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of the 15th International Conference on Multimedia*. ACM Press, 991–1000.

NAVIGLI, R. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv. 41*, 2, 1–69.

NAVIGLI, R. AND VELARDI, P. 2003. An analysis of ontology-based query expansion strategies. In *Proceedings of the ECML/PKDD-2003 Workshop on Adaptive Text Extraction and Mining*.

NAVIGLI, R. AND VELARDI, P. 2005. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell. 27*, 7, 1075–1086.

OSIŃSKI, S. AND WEISS, D. 2005. A concept-driven algorithm for clustering search results. *IEEE Intell. Syst. 20*, 3, 48–54.

PALLETI, P., KARNICK, H., AND MITRA, P. 2007. Personalizedweb search using probabilistic query expansion. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, 83–86.

PARK, L. A. F. AND RAMAMOHANARAO, K. 2007. Query expansion using a collection dependent probabilistic latent semantic thesaurus. In *Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'07)*. Springer, 224–235.

PERUGINI, S. AND RAMAKRISHNAN, N. 2006. Interacting with web hierarchies. *IT Professional 8*, 4, 19–28.

PIRKOLA, A., HEDLUND, T., KESKUSALO, H., AND JĂRVELIN, K. 2001. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Info. Retrieval 4*, 209–230.

PORTER, M. F. 1982. Implementing a probabilistic information retrieval system. *Info. Technol.: Resear. Develop. 1*, 2, 131–156.

PORTER, M. F. 1997. An algorithm for suffix stripping. In *Readings in Information Retrieval*, K. S. Jones and P. Willett Eds., Morgan Kaufmann, 313–316.

QIU, Y. AND FREI, H.-P. 1993. Concept-based query expansion. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 160–169.

RIEZLER, S., VASSERMAN, A., TSOCHANTARIDIS, I., MITTAL, V., AND LIU, Y. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*. Association for Computational Linguistics, 464–471.

ROBERTSON, S. E. 1986. On relevance weight estimation and query expansion. *J. Documentation 42*, 3, 182–188.

ROBERTSON, S. E. 1990. On term selection for query expansion. *J. Documentation 46*, 4, 359–364.

ROBERTSON, S. E. AND SPARCK JONES, K. 1976. Relevance weighting of search terms. *J. Amer. Soc. Info. Science 27*, 129–146.

ROBERTSON, S. E. AND WALKER, S. 2000. Microsoft cambridge at trec-9: Filtering track. In *Proceedings of the 9th Text REtrieval Conference (TREC-9)*. NIST Special Publication 500-249. National Institute of Standards and Technology (NIST), Gaithersburg, MD, 361–368.

ROBERTSON, S. E., WALKER, S., AND BEAULIEU, M. M. 1998. Okapi at TREC-7: Automatic ad hoc, filtering, VLC, and interactive track. In *Proceedings of the 7th Text REtrieval Conference (TREC-7), NIST Special Publication 500-242*. National Institute of Standards and Technology (NIST), Gaithersburg, MD, 253–264.

ROCCHIO, J. J. 1971. Relevance feedback in information retrieval. In *The SMART Retrieval System*, G. Salton Ed., Prentice-Hall, Englewood Cliffs, NJ, 313–323.

RUTHVEN, I. 2003. Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 213–220.

RUTHVEN, I. AND LALMAS, M. 2003. A survey on the use of relevance feedback for information access systems. *Knowl. Engin. Rev. 18*, 2, 95–145.

SAHLGREN, M. 2005. An introduction to random indexing. In *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*.

SAKAI, T., MANABE, M., AND KOYAMA, M. 2005. Flexible pseudo-relevance feedback via selective sampling. *ACM Trans. Info. Syst. 4*, 2, 111–35.

SALTON, G. AND BUCKLEY, C. 1990. Improving retrieval performance by relevance feedback. *J. Amer. Soc. Info. Science 41*, 4, 288–297.

SALTON, G. AND MCGILL, M. 1983. *Introduction to Modern Information Retrieval*. McGraw Hill, New York, NY.

SANDERSON, M. 1994. Word sense disambiguation and information retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 142–151.

SANDERSON, M. 2000. Retrieving with good sense. *Info. Retrieval 2*, 1, 49–69.

SAVOY, J. 2005. Comparative study of monolingual and multilingual search models for use with asian languages. *ACM Trans. Asian Lang. Info. Process. 4*, 2, 163–189.

SCHLAEFER, N., KO, J., BETTERIDGE, J., SAUTTER, G., AND AMD E. NYBERG, M. P. 2007. Semantic extensions of the Ephyra QA system for TREC 2007. In *Proceedings of the 16th Text REtrieval Conference (TREC'07)*. NIST Special Publication 500-274. National Institute of Standards and Technology (NIST), Gaithersburg, MD, 332–341.

SCHÜTZE, H. AND PEDERSEN, J. O. 1995. Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*. 161–175.

SCHÜTZE, H. AND PEDERSEN, O. 1997. A co-occurrence based thesaurus and two applications to information retrieval. *Info. Process. Manage. 33*, 3, 307–318.

SEMERARO, G., LOPS, P., BASILE, P., AND DE GEMMIS, M. 2009. On the tip of my thought: Playing the guillotine game. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*. AAAI Press, 1543–1548.

SHEN, X. AND ZHAI, C. 2005. Active feedback in ad hoc information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 59–66.

SHOKOUHI, M., AZZOPARDI, L., AND THOMAS, P. 2009. Effective query expansion for federated search. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 427–434.

SINGHAL, A. AND PEREIRA, F. 1999. Document expansion for speech retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 34–41.

SONG, M., SONG, I.-Y., ALLEN, R. B., AND OBRADOVIC, Z. 2006. Keyphrase extraction-based query expansion in digital libraries. In *Proceedings of the 6th ACM/IEEE-CS joint International Conference on Digital Libraries (JCDL'06)*. ACM Press, 202–209.

SONG, M., SONG, I.-Y., HU, X., AND ALLEN, R. B. 2007. Integration of association rules and ontologies for semantic query expansion. *Data Knowl. Engin. 63*, 1, 63–75.

SUN, R., ONG, C.-H., AND CHUA, T.-S. 2006. Mining dependency relations for query expansion in passage retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 382–389.

SURYANTO, M. A., LIM, E.-P., SUN, A., AND CHIANG, R. H. 2007. Document expansion versus query expansion for ad-hoc retrieval. In *Proceedings of the ACM 1st Workshop on CyberInfrastructure: Information Management in eScience*. ACM Press, 47–54.

THEOBALD, M., SHENKEL, R., AND WEIKUM, G. 2004. Top-k query evaluation with probabilistic guarantees. In *Proceedings of the 13th International Conference on Very Large Data Bases*. ACM Press, 648–659.

THEOBALD, M., SHENKEL, R., AND WEIKUM, G. 2005. Efficient and selftuning incremental query expansion for top-k query processing. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 242–249.

VAN RIJSBERGEN, C. J. 1979. *Information Retrieval*. Butterworths.

VECHTOMOVA, O. 2009. Query expansion for information retrieval. In *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu Eds., Springer, 2254–2257.

VECHTOMOVA, O. AND KARAMUFTUOGLU, M. 2004. Elicitation and use of relevance feedback information. *Info. Process. Manage. 42*, 1, 191–206.

VÉRONIS, J. 2004. HyperLex: lexical cartography for information retrieval. *Computer Speech Lang. 18*, 3, 223–252.

VOORHEES, E. 1993. Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 171–180.

VOORHEES, E. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 61–69.

VOORHEES, E. 2004. Overview of the trec 2004 robust track. In *Proceedings of the 13th Text REtrieval Conference (TREC-7)*. NIST Special Publication 500-261. National Institute of Standards and Technology (NIST), Gaithersburg, MD.

VOORHEES, E. AND HARMAN, D. 1998. Overview of the seventh text retrieval conference (TREC-7). In *Proceedings of the 7th Text REtrieval Conference (TREC-7)*. NIST Special Publication 500-242. National Institute of Standards and Technology (NIST), Gaithersburg, MD, 1–24.

WANG, H., LIANG, Y., FU, L., XUE, G.-R., AND YU, Y. 2009. Efficient query expansion for advertisement search. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 51–58.

WANG, X., FANG, H., AND ZHAI, C. 2008. A study of methods for negative relevance feedback. In *Proceedings of the 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 219–226.

WEI, X. AND CROFT, W. B. 2007. Modeling term associations for ad-hoc retrieval performance within language modeling framework. In *Proceedings of the 29th European Conference on IR Research (ECIR'07)*. Springer, 52–63.

WHITE, R. W., RUTHVEN, I., AND JOSE, J. M. 2005. A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 35–42.

WINAVER, M., KURLAND, O., AND DOMSHLAK, C. 2007. Towards robust query expansion: Model selection in the language modeling framework. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 729–730.

WITTEN, I. H., MOFFAT, A., AND BELL, T. C. 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images* 2nd Ed. Morgan Kaufman.

WONG, S. K. M., ZIARKO, W., RAGHAVAN, V. V., AND WONG, P. C. N. 1987. On modeling of information retrieval concepts in vector spaces. *ACM Trans. Datab. Syst. 12*, 2, 299–321.

WONG, W. S., LUK, R. W. P., LEONG, H. V., HO, K. S., AND LEE, D. L. 2008. Re-examining the effects of adding relevance information in a relevance feedback environment. *Info. Process. Manage. 44*, 3, 1086–1116.

XU, J. AND CROFT, W. B. 1996. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 4–11.

XU, J. AND CROFT, W. B. 2000. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Info. Syst. 18*, 1, 79–112.

XU, Y., JONES, G. J. F., AND WANG, B. 2009. Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 59–66.

XU, Z. AND AKELLA, R. 2007. Incorporating diversity and density in active learning for relevance feedback. In *Proceedings of the 29th European Conference on IR Research (ECIR'07)*. Springer, 246–257.

XUE, G.-R., ZENG, H.-J., CHEN, Z., YU, Y., MA, W.-Y., XI, W., AND FAN, W. 2004. Optimizing web search using web click-through data. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*. ACM Press, 118–126.

YIN, Z., SHOKOUHI, M., AND CRASWELL, N. 2009. Query expansion using external evidence. In *Proceedings of the 31th European Conference on Information Retrieval (ECIR'09)*. Springer, 362–374.

YU, S., CAI, D., WEN, J. R., AND MA, W. Y. 2003. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *Proceedings of the 12th International Conference on World Wide Web*. ACM, 11–18.

ZELIKOVITZ, S. AND HIRSH, H. 2000. Improving short-text classification using unlabeled background knowledge to assess document similarity. In *Proceedings of the 17th International Conference on Machine Learning (ICML'00)*. National Institute of Standards and Technology (NIST), 1183–1190.

ZHA, Z.-J., YANG, L., MEI, T., WANG, M., AND WANG, Z. 2009. Visual query suggestion. In *Proceedings of the 17th ACM International Conference on Multimedia*. ACM Press, 15–24.

ZHAI, C. AND LAFFERTY, J. 2001a. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the 10th International Conference on Information and Knowledge Management*. ACM Press, 403–410.

ZHAI, C. AND LAFFERTY, J. 2001b. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 334–342.

ZIMMER, C., TRYFONOPOULOS, C., AND WEIKUM, G. 2008. Exploiting correlated keywords to improve approximate information filtering. In *Proceedings of the 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 323–330.

Received November 2009; revised February 2010; accepted March 2010