SMERA: Semantic Mixed Approach for Web Query Expansion and Reformulation

Bissan Audeh, Philippe Beaune, Michel Beigbeder

SMERA: Semantic Mixed Approach for Web Query Expansion and Reformulation

Bissan Audeh, Philippe Beaune and Michel Beigbeder

Abstract Matching users' information needs and relevant documents is the basic goal of information retrieval systems. However, relevant documents do not necessarily contain the same terms as the ones in users' queries. In this paper, we use semantics to better express users' queries. Furthermore, we distinguish between two types of concepts: those extracted from a set of pseudo relevance documents, and those extracted from a semantic resource such as an ontology. With this distinction in mind we propose a Semantic Mixed query Expansion and Reformulation Approach (SMERA) that uses these two types of concepts to improve web queries. This approach considers several challenges such as the selective choice of expansion terms, the treatment of named entities, and the reformulation of the query in a user-friendly way. We evaluate SMERA on four standard web collections from INEX and TREC evaluation campaigns. Our experiments show that SMERA improves the performance of an information retrieval system compared to non-modified original queries. In addition, our approach provides a statistically significant improvement in precision over a competitive query expansion method while generating concept-based queries that are more comprehensive and easy to interpret.

1 Introduction

Once the domain of librarians and specialists, today the practice of searching for information is open to users from different profiles and backgrounds, all of whom use queries composed of keywords to look for information on the web. The challenge of this online search for content is that retrieval systems need to provide relevant documents for all the users who express the need for a particular piece of

B. Audeh (✉) · P. Beaune · M. Beigbeder
École Nationale Supérieure des Mines de Saint-Étienne,
158 Cours Fauriel, 42023 Saint-Étienne, France
e-mail: audeh@emse.fr

P. Beaune
e-mail: beaune@emse.fr

information using many different queries. In addition, the length of web queries is a major challenge for most query modification approaches.

The issue we are tackling is how to improve the precision of short ambiguous web queries. To achieve this goal, our paper explores semantic related techniques for automatic query reformulation.

Since most web users employ two to three terms in a query to express their information needs (Jansen et al. 2000), it is not easy for a system to retrieve relevant documents at early ranks in the result list. To address this challenge, a number of approaches propose to consider the semantics during the indexing step. In this case, concepts, instead of terms (or stems), are used to index documents and queries. The relevance between a document and a query is then evaluated on the basis of this conceptual indexation. Another option is to keep a keyword-based index and to use semantic approaches to expand and reformulate users' queries. While both of these solutions have been explored in the literature of information retrieval, in general, it is not possible to confirm the advantage of one option over the other one. Many elements could affect the choice of how to use the semantics within an information retrieval system, such as the nature of the document collection (web, closed collection), the context of use (professional, general), the motivation (creating a new retrieval system or improving an existing one), and the cost. In this paper, we are interested in the case where documents and queries are indexed using classical term-based techniques. Thus, we focus on semantically modifying users' queries while preserving the keyword-based retrieval mechanism.

Techniques that automatically modify users' queries have existed since the early years of information retrieval. As a result, the literature is wealthy of terms like "query expansion", "query refinement", "query reformulation", "query enrichment", "local and global analysis" and "relevance feedback". All these techniques intend to improve keyword-based queries even though the number of terms used to describe how this is achieved is confusing. For our work, we employ two commonly used terms: *query expansion* and *query reformulation*. We define *query expansion* as assigning new terms to users' queries, whereas we consider *query reformulation* as the way in which these new terms are integrated within the original query. The literature does not always make a difference between *query expansion* and *query reformulation*, this is because in most cases the query is considered as a bag of words. In general, approaches try to add new terms with eventually optimized weights; hence, reformulating the query is not considered as a separate process.

In this paper, we study the effect of different semantic aspects to automatically improve web queries. To do this, we associate query terms with implicit concepts that we obtain with a pseudo relevance feedback approach, and explicit concepts that we extract from an ontology. Once detected, explicit and implicit concepts are used to obtain sets of expansion terms (Sect. 3.1) and to construct a new query (Sect. 3.2). The new query is still composed of keywords, but it is structured so as to represent the concepts. This allows a straightforward understanding of the relationships between the original user keywords and the detected concepts. In Sect. 4 we compare our proposition versus no query expansion as well as versus a state-of-the-art expansion approach. We begin our paper with a brief state of the art of existing query expansion and query reformulation approaches.

2 Query Expansion and Reformulation in Information Retrieval

Associating new terms to a query requires the use of a data source other than the query itself. This resource can be a collection of documents (Qiu and Frei 1993), a subset of the collection via a relevance or pseudo relevance feedback process (Rocchio and Salton 1965), a completely independent resource that is also a collection of documents (Deveaud et al. 2013), or a semantic resource (Voorhees 1994). All of these approaches have been the subject of many comparisons and surveys that as a whole reveal three common points: an expanded query is often not structured, named entities are processed in the same way as common terms, and no specific consideration is taken regarding the advantage (or disadvantage) of adding a candidate term to the query. In the following subsections, we will focus on query expansion or reformulation approaches that consider these three aspects in the state of the art.

2.1 Concept-Based Query Reformulation

Representing a query completely depends on the query language that the retrieval system can interpret. A bag-of-words representation is the most common way to reformulate an expanded query. With this representation, the query is composed of weighted terms with no explicit operators.

In the literature, several approaches explored the advantages of structured queries, whether by using only original query terms (the case of studies on long queries) (Metzler and Croft 2005; Bendersky and Croft 2008; Maxwell and Croft 2013), or by integrating new terms from different resources with the original query terms (Bendersky et al. 2011, 2012; Deveaud et al. 2013). Query expansion approaches, in the latter case, propose to introduce the notion of concepts into the expanded query, which we call "concept based query representation". For (Bendersky et al. 2011), a concept is one or more terms that must belong to one of the following types: an original query term, a composition of multiple original terms, or term obtained from the pseudo relevance feedback of the original query on different expansion collections. The obtained concepts are then combined to construct a new query using Eq. 1:

$$Score(Q, D) = \sum_{T \in \tau} \sum_{\kappa \in T} \lambda_\kappa f(\kappa, D) \tag{1}$$

where τ is the set of concept types, $f(\kappa, D)$ is the query likelihood retrieval function that matches the concept κ in the document D, and λ_κ is the weight of the concept κ. The weight in this equation takes a set of features into account, especially the frequency of the concept in the expansion collections. Similarly, (Deveaud et al. 2013) work on detecting query concepts but without considering possible associations among original terms. So, a concept in this case is either an original query

term or a set of terms from pseudo relevance feedback documents. (Deveaud et al. 2013) use Latent Dirichlet Allocation (LDA) (Blei et al. 2003) on the document sets obtained by pseudo relevance feedback on different collections. The score of a document is computed as shown by Eq. 2:

$$Score(Q, D) = \lambda \cdot P(Q|D) + (1 - \lambda) \cdot \prod_{k \in T_{\hat{K}}} \hat{\delta}_k \prod_{w \in W_k} \hat{\phi}_{k,w} \cdot P(w|D) \qquad (2)$$

where W_k is the set of terms of the concept k, $\hat{\phi}_{k,w}$ is the weight of the term w in the concept k, $\hat{\delta}_k$ is the normalized weight of the concept k, and $T_{\hat{K}}$ is the set of concepts assigned to the query. The authors show that combining four different collections for concept extraction is more effective in precision than the use of any single resource.

All of these approaches did generate structured queries based on the notion of concepts, but they didn't explore the advantage of using formal semantic relationships from a structured resource like an ontology. They also did not consider the specificity of named entities.

2.2 Query Expansion and Named Entities

The approaches in the previous subsection focused on pseudo relevance feedback techniques, where named entities are not considered as special terms. For this reason, it is possible that the expanded query doesn't contain any reference to these important objects. Other approaches focus on the importance of named entities in a query; for example, studies on long queries consider a sub-query containing a named entity as a valuable reformulation candidate (Huston and Croft 2010; Kumaran and Carvalho 2009). Bendersky and Croft (2008) classify noun phrases (eventually named entities) in order to use them in the reformulated query. Recent approaches are becoming increasingly interested in Wikipedia, which is a rich resource of named entities. Xu et al. (2008) extracted terms from Wikipedia, that are semantically close to named entities in the query, while Brandao et al. (2011) proposed an approach based on the infoboxes of Wikipedia to expand named entities.

These approaches explicitly handle named entities differently from other terms. Nevertheless, they rely on a bag-of-words representation for the modified query instead of concept-based representation. In addition, no specific treatment is done to control the quality of expansion terms.

2.3 Quality of Expansion Terms

For most query expansion approaches, new terms are systematically added to all queries, even if in some instances, better results can be obtained without expansion. These approaches do not consider the advantage of adding (or not adding) each term

to the query. Though, several methods exist to measure the quality of a query or query terms that could be used in query expansion, (Cronen-Townsend et al. 2002) proposed the clarity measure, which is based on computing the entropy between the query model and the document model. They confirmed the relationship between this measure and query ambiguity. Nevertheless, the effectiveness of using this measure to choose new terms for query expansion was not confirmed (Zobel 2004; Shah and Croft 2004). Other studies focused on measuring the importance of the query terms. (Zhao and Callan 2010) used a technique based on pseudo relevance feedback and latent semantic analysis (Deerwester et al. 1990) to classify terms according to their importance within the query. This approach was only used to evaluate original query terms, not to choose new terms for query expansion purpose.

From this brief presentation of query modification approaches, it can be seen that structured queries, named entities and terms quality aspects are the subject of several studies in the dedicated literature. We consider that an approach that gathers all of these aspects could be effective to improve web queries. For this reason, we propose the semantic mixed expansion and reformulation approach that we thoroughly discuss in the following section.

3 Semantic Mixed Expansion and Reformulation Approach (SMERA)

As discussed in Sect. 2, query expansion and reformulation approaches are not new to information retrieval. In light of the weaknesses revealed by these approaches, we propose SMERA that uses semantics to improve web queries. Our approach utilizes both, but well distinguished, query expansion and query reformulation techniques (Fig. 1). The consideration of concept-based query representation, named entities and the quality of expansion terms allows our approach to generate queries that are comprehensible and easy to interpret. We believe that generating user-friendly queries is important to understand and analyze the relationships between a well-expressed query (from a human point of view) and an effective query[1].

As Fig. 1 shows, our approach is composed of two steps. The first step (expansion) includes detecting query concepts, and choosing the most appropriate and representative terms of these concepts. The second step (reformulation) will use the concepts detected during the first step to reformulate the expanded query in a concept-based representation.

[1] In this paper we define an effective query is the one that obtains good results with standard measures used in evaluation campaigns, in particular, precision measures for the case of web queries.

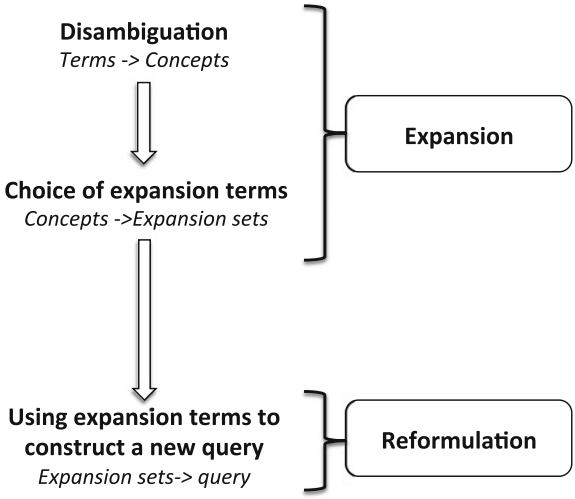

Fig. 1 The main steps of SMERA

3.1 Expansion

Defining query expansion as a separate step that assigns new terms to the query allows it to be independent from the matching function of the retrieval model. Our approach depends on the assumption that each original query term (except stop words) belongs to a concept (in its abstract meaning). This strong assumption is justified by the fact that a user doesn't use one term to express two different concepts; on the contrary, he may use multiple terms to express one concept. Thus, we consider that a query of k terms corresponds to at most k concepts. This allows us to initialize the number of concepts and to keep a clear relationship between original terms and represented concepts in the reformulated query. For each query term, we define an expansion set that contains semantically similar terms. The nature of web queries (short, ambiguous, and rich with named entities) and the literature of query expansion, oriented our approach towards mixing two types of expansion resources: the collection of documents through pseudo relevance feedback and an ontology. We use pseudo relevance feedback documents to detect what we call "implicit concepts", while we consider named entities in the query as the "explicit concepts" that we identify using an ontology. In both cases, a concept in our approach is a set of semantically similar terms.

Figure 2 shows the main steps of our expansion approaches for both named entities and other terms. If we consider the example in Table 1, SMERA will first detect the named entity "Jack Robinson". This named entity will be disambiguated and linked to an explicit concept which is then expanded with the ontology-based approach (cf. Sect. 3.1.1). The other terms of the query, except stop words, will be expanded based on implicit concepts extracted by an LSI-based[2] method (cf. Sect. 3.1.2).

[2]LSI: Latent Semantic Indexing (Deerwester et al. 1990).

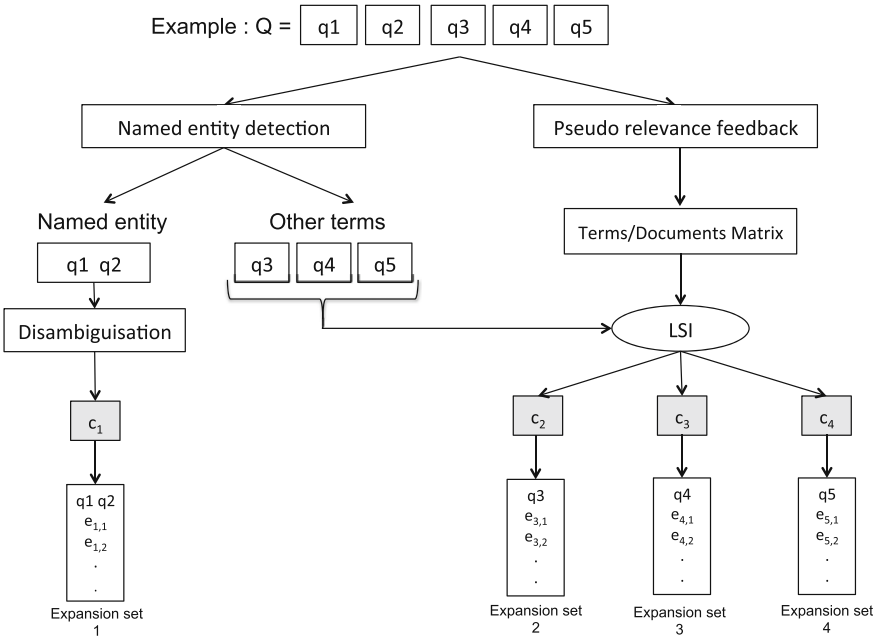

Fig. 2 Generating expansion sets in SMERA from explicit and implicit concepts

Table 1 Term categories and SMERA actions demonstrated on one TREC query (#455)

Original query: When did Jack Robinson appear at his first game?		
Category	Values	SMERA action
Named entities	"Jack Robinson"	Expand using explicit concept approach
Non named entity term	Appear, first, game	Expand using implicit concept approach
Stop words	When, did, at, his	Do not expand

Expansion sets do not necessarily have the same size. This is because the number of available terms in the corresponding concepts is not necessarily the same. In addition, we use quality filters that measure the utility of adding a term to the query and eliminate less useful terms. As a result, our approach does not always expand all queries with the same number of expansion terms. A detailed explanation of these steps is in the following subsections.

3.1.1 Ontology-Based Approach

The role of an ontology in our expansion approach is first to provide a semantic resource in which named entities can be identified as concepts. Once identified in the ontology, semantic relationships could be of use to reach the appropriate expansion terms. In a web context, a single domain ontology can not be used for all queries. The generic semantic resource most commonly used in information retrieval is WordNet (Miller et al. 1990). However, this resource's main problem, in our case, is its lack of named entities. To overcome this issue, we sought yet another alternative: the ontology YAGO (Suchanek et al. 2007). The advantage of this ontology is that it gathers WordNet and Wikipedia, inheriting the formally organized structure of the former and the supply of named entities of the latter, which makes it suitable for our named entity expansion.

To find its expansion set, a named entity has to be identified in YAGO. For this purpose, we use the disambiguation approach of Aida (Hoffart et al. 2011). This approach selects all possibly corresponding concepts in YAGO for each named entity in a query and calculates disambiguation scores for these candidate concepts. The concept that obtains the highest score is considered to be the one corresponding to the named entity in the query. Concepts obtained using this approach are considered by SMERA as explicit concepts.

A wealthy number of semantic relationships exist in YAGO. For example, in the case of concepts corresponding to a named entity, we can find relationships like "lives in" for person entities, or "has the surface" if the named entity is a city. On the other hand, all named entity concepts in YAGO have the semantic relationship "rdf:label". This relationship corresponds to the "redirect" link in Wikipedia, it links the named entity to all its possible appellations. These appellations can be simply orthographic alternative names (e.g., Baltimore-Baltamore), syntactically different names (e.g., Baltimore-Mobtown), or even nominal phrases (e.g., "Aleck Bell"-"The father of the deaf"). In this work, we choose the relationship (rdf:label) to expand named entities. This choice assumes that using alternative appellations to expand named entities leads to less query drift risk than using other semantic relationships in YAGO. In our previous example of Table 1, after disambiguation, the named entity "Jack Robinson" obtains two expansion terms: "Jackie Robinson" and "Jack Roosevelt Robinson".

3.1.2 Pseudo Feedback Approach

The idea of this approach is to detect implicit concepts from a set of pseudo feedback documents related to users' initial query. Several methods exist to extract concepts from a set of documents, such as LDA (Blei et al. 2003), ESA (Gabrilovich and Markovitch 2007) or LSI (Deerwester et al. 1990). We chose to use LSI because of its ability to detect high-level co-occurrence relationships between terms. In other words, two terms that do not occur together in a studied set of documents, but do frequently co-occur with a third term, are considered by LSI as semantically related. To achieve this, LSI (Deerwester et al. 1990) starts by applying singular

value decomposition on a matrix A of m lines (m terms) and n columns (n feedback documents), which contains frequencies tf of the terms in document collection (in our case, pseudo feedback documents). The results of this step are the three matrices presented in Eq. 3,

$$A_{\{m,n\}} = U_{\{m,m\}} S_{\{m,n\}} V^T_{\{n,n\}} \tag{3}$$

where S is the diagonal matrix that contains singular values of A. The theory of LSI is that reducing the dimension of the three resulting matrices gives an approximation of the original matrix A and reduces the noise (Eq. 4).

$$A'_{\{m,n\}} = U_{\{m,k\}} S_{\{k,k\}} V^T_{\{k,n\}} \tag{4}$$

In our pseudo feedback expansion approach, we are interested in the matrix $U_{\{m,k\}}$. This matrix contains the m vectors of terms appearing in pseudo relevance feedback documents. These vectors belong to the semantic space of k dimensions created by LSI (Fig. 3).

To find the expansion set of a query term q, we measure its similarity with a term that appears in the feedback documents by calculating its distance with this term[3]. We then suppose that the terms that are the most similar to q belong to the same implicit concept, as presented in Fig. 3. In some cases, an expansion term q' of a term q is also a query term; in this case, we consider that both terms q and q' belong to the same implicit concept (c2 in Fig. 4) and they will both correspond to one expansion set in the reformulated query.

In our example of Table 1, two expansion sets were found for three non-named entity terms: {appear}, {first, play, team, season, game, ball}. From these two sets, we can see that the implicit concepts related to the query terms "first" and "game" were merged resulting in one expansion set for both of these terms.

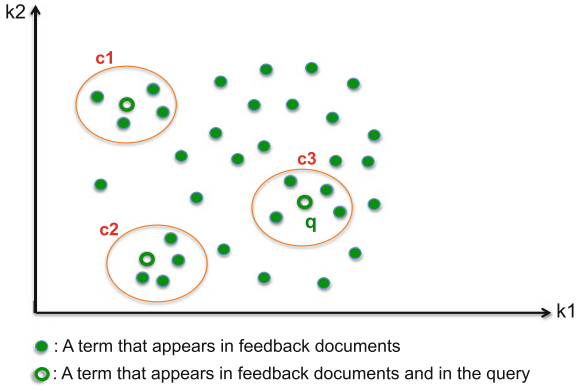

Fig. 3 Terms of feedback documents in the semantic space of LSI (example for the case of 2 dimensions k1 et k2)

[3]Our experiments showed no significant difference between using euclidian and cosine distances, in this paper we used euclidian distance because it is more clear for our graphical demonstration in Figs. 3 and 4.

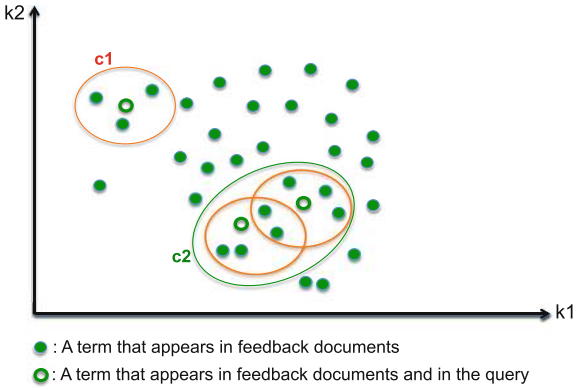

Fig. 4 The fusion of expansion sets in the case of query terms that are semantically close in LSI semantic space

3.1.3 Quality of Expansion Terms

Our ontology-based and feedback-based approaches presented in Sects. 3.1.1 and 3.1.2 respectively generate expansion sets for, at most, each original query term. In this work, we consider the quality of terms, which means their usefulness in obtaining relevant documents to the user's information need. From this point of view, we consider original query terms as the most valuable terms in the query because they were chosen by the user to express his own information need. An expansion term, on the other hand, is considered to be useful if it is not too generic, and as far as we are sure it belongs to a valid query concept. To express these two subjective conditions, we define *specificity* and *certitude* qualities. *Specificity* is a boolean value. *Certitude* is measure with values between 0 and 1. Expansion terms that do not satisfy a minimum threshold for this measure are rejected from their expansion sets and will not appear in the reformulated query.

Concerning the specificity, we consider named entities as specific terms. For non named entity terms, since the use of verbs and adverbs in web queries is not frequent (Barr et al. 2008), we only compute specificity for nouns. For this purpose, we use the taxonomy of WordNet, whereby generic terms are placed in the top of the hierarchy while specific terms can be found in deeper levels. Thus a noun is added to an expansion set if its depth is greater than a threshold.

The certitude is directly related to the process that links a query term to its corresponding implicit or explicit concept. For the feedback approach, the choice of an expansion term depends of its semantic similarity, in the LSI space, with the original query term. Hence, a term that is semantically closer to the original term is likely a more suitable expansion term. In this case, we define the certitude score between a term t and a query term q as the euclidean distance between their corresponding vectors ($\overrightarrow{t}$, $\overrightarrow{q}$) in the LSI space as defined in Eq. 5.

$$Cert(t, q) = Dist_{euclide}(\overrightarrow{t}, \overrightarrow{q}) \tag{5}$$

As mentioned in Sect. 3.1.1, an explicit concept is chosen in YAGO for a named entity in the query if it obtains the maximum disambiguation score. For each expansion term that we obtain by the relation "rdf:label", the certitude value is the disambiguation score S_{dis} of the concept to which the query term belongs (Eq. 6)

$$Cert(t, q) = S_{dis}(q, c) \tag{6}$$

where q is a query term, c is the disambiguated YAGO concept associated to q, t is a possible expansion term associated to the concept c, and $S_{dis}(q, c)$ is the disambiguation score, obtained by Aida (Hoffart et al. 2011), for the query term q and the concept c.

3.2 Concept-Based Query Reformulation

Up to this point, the expansion approaches we proposed are independent of the retrieval model. However, reformulating a query depends on the retrieval model and its query language. To achieve a concept-based query representation, we need a structured query language that supports three main elements: proximity between terms, synonymy and term weighting. The model proposed by (Metzler and Croft 2004) is a good environment to apply our idea of semantic reformulation. In the next subsections, we present an overview of this model and how we use it query language to reformulate users' query.

3.2.1 The Retrieval Model of Metzler and Croft (2004)

This information retrieval model is a combination of inference networks and query likelihood models. Like in inference network models, it is possible to handle structured queries, but estimating the probabilities is achieved using a query likelihood language model. The model is implemented within the framework Indri (Strohman et al. 2004), which is part of the Lemur[4] project. Indri proposes a query language model that allows expressing the different functionality of the retrieval model. Table 2 represents some demonstrative examples cited in the Lemur wikipage[5] and shows how the implementation of (Metzler and Croft 2004) in Indri handles the different query language operators.

[4] http://sourceforge.net/p/lemur/wiki/The%20Indri%20Query%20Language.

[5] http://sourceforge.net/p/lemur/wiki/Belief%20Operations/.

Table 2 Demonstrative examples of the functionality of Indris operators

Syntax	Interpretation
#combine(dog train)	$0.5log(b(dog)) + 0.5log(b(train))$
#weight(1.0 dog 0.5 train)	$0.67log(b(dog)) + 0.33log(b(train))$
#wsum(1.0 dog 0.5 dog)	$log(0.67b(dog) + 0.33b(dog))$
#syn(car automobile)	one occurrence of "car" or "automobile"
#wsyn(1.0 car 0.5 automobile)	like #syn, but the occurrence of "car" counts as twice the occurrence of "blue"
#n(blue car)	"blue" appears before "car" in a window of maximum n words
#uwn(blue car)	"blue" appears before or after "car" in a window of maximum n words

3.2.2 Representing Concepts in Keyword Query

Our reformulation approach considers the final query to be a linear combination of the user's original query and the combination of the different expansion sets according to three aspects: proximity, synonymy and weighting. The score of this reformulated query is calculated with Eq. 7

$$p(Q|d) = \lambda \prod_{q} p(q|d) + (1 - \lambda) \prod_{i=1}^{k} b(r_i)^{w_i} \tag{7}$$

where $p(q|d)$ is the query likelihood probability for the original query term q and a document d, r_i is the combination of terms of an expansion set with an Indri operator (#combine, #weight or #syn), and $b(r_i)$ is the belief calculated according to (Metzler and Croft 2004) as illustrated in Table 2. Finally w_i is the weight of the estimated belief of the representation r_i. In this current study, expansion sets are considered to be equally important to the query ($w_i = 1$, for all i).

For example, the reformulation of the query presented previously in Table 1 is demonstrated in Fig. 5. This figure shows how we combine expansion sets using

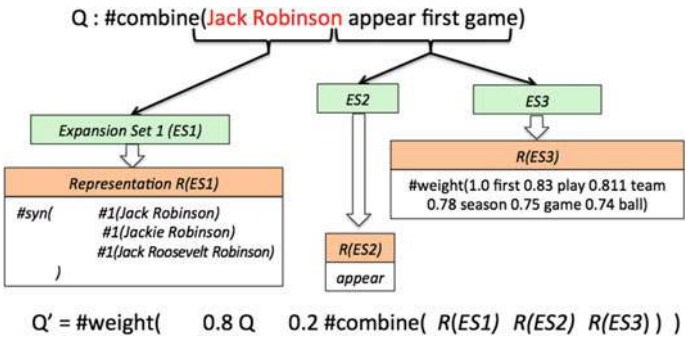

Fig. 5 Example of query reformulation using SMERA

Eq. 7 and the corresponding operators in Indri model (cf. Table 2). The following subsections explain why and how each operator is used to reformulate a query in our approach.

Proximity

Expansion terms that come from the ontology YAGO can be expressions or names composed of one or multiple terms, as we have seen in Sect. 3.1.1. When an element of a named entity expansion set is composed of multiple terms, the proximity between these terms should be highly respected while representing this entity in a query. Expressing the proximity between terms in the query implies defining the maximum distance within which these terms could be considered as related to the entity. In addition, we have to precise if the order in which these terms appear in a text is important. In our work, we suppose that the coverage of semantic alternatives of named entities is the responsibility of the resource, which in our case is the YAGO ontology. For this reason, we consider expansion elements obtained from the semantic resource as blocks that should appear verbatim in a relevant document. Thus, our approach requires that terms that belong to the same expression should be within a window of width 1 and appear in the exact order as in the semantic resource. To represent these types of expansion terms we use the operator #1 (cf. Table 2 and Fig. 5).

Synonymy

Our expansion approaches extract terms that are semantically related to query terms. This semantic similarity is not the direct synonymy in the case of our feedback approach, and we do not use this functionality for feedback expansion sets. In the case of named entities expansion, the semantic similarity is defined by the explicit relation "rdf:label", which will give possible, semantically equal, alternatives of the named entity. When evaluating a document that contains one of these alternatives, we want the matching function of the retrieval model to consider it as any of its other alternatives. For this reason, expansion sets that are obtained from an explicit concept are represented by the operator #syn in the Indri's query language. It should be noted that expansion terms of a named entity are not weighted in our current approach, we consider them as equally important synonyms, though exploring weighting possibility based on popularity or corpus statistics is an interesting area for future work.

Weighting

In our reformulation approach, weighting a term means defining its importance in its expansion set. We consider original query terms as important (weight = 1). The more an expansion term is close to an original query term, the more its weight is close

to 1. As we mentioned in the synonymy section, this notion is not defined when the expansion set is obtained from YAGO because we consider all its terms as equal. On the other hand, expansion terms obtained by the feedback approach are terms that are statistically close to the original term in the LSI space, but they cannot be considered as synonyms. In this paper we explore the effect of using the similarity distance from the query term as a weight in the reformulated query. Expansion terms that are obtained from the feedback approach are combined with the operator #weight in the Indri's query language. The euclidean distance between an expansion term and its original term (cf. Eq. 5) is considered as its weight in the #weight expression. In the example of Fig. 5, the original query term "first" has the weight 1, while expansion terms have decreasing weights according to their semantic similarity with this term.

4 Experiments and Evaluation

4.1 Framework

To evaluate our semantic mixed expansion and reformulation approach (SMERA), we used four web collections from TREC and INEX evaluation campaigns, as displayed in Table 3. All of these collections were indexed with the same parameters using Indri: standard stop words were removed and a Krovetz Stemmer was used.

As a baseline, we used the query likelihood language model (Ponte and Croft 1998) to run the users' queries without expansion; we called this the QL model. We also used the relevance model approach (RM3) (Lavrenko and Croft 2001) as a reference model for query expansion. Both QL and RM3 are implemented in the Indri's framework. In addition to these reference approaches, we compared SMERA to the use of only one method for query expansion: the use of LSI via pseudo relevance feedback to expand query terms (both common terms and named entities), we called this the LSI approach, and the use of YAGO to disambiguate and expand named entities (the YAGO approach). The evaluation measures that we used in this experience are precision measures (MRR, P@10 and MAP), which are the most important in our

Table 3 Information about the queries used in our experiments

	# documents	queries	year (track)	nb. judged queries	nb. named entities
Inex 2006	659, 388	544–677	2008 (ad hoc)	70	23
Inex 2009	2, 666, 190	1–115	2009 (ad hoc)	68	21
WT10g	1, 692, 096	451–550	2000–2001 (Web ad hoc)	98	25
Gov2	25, 205, 179	701–850	2004–2006 (Terrabyte)	148	47

Table 4 Free parameters for all the approaches of our experiments

Parameter	Description	Approach
μ	Dirichlet smoothing	QL, SMERA, RM3
n_{Smera}, n_{Rm3}	Number of feedback documents	SMERA, RM3
t	Number of expansion terms	RM3
m	Number of expansion terms per concept	SMERA
k	Number of LSI dimensions	SMERA
$\alpha 1$	The threshold of the certitude filter	SMERA
$\alpha 2$	The depth threshold of the specificity filter	SMERA
λ_{Smera}, λ_{Rm3}	The weight of the original part against the expanded part of the reformulated query (Eq. 7)	SMERA, RM3

web context, in addition to ROM (Audeh et al 2013), which is a Recall Oriented Measure that also takes precision into account.

An interesting aspect of our approach is the scalability. In fact, SMERA applies LSI to a small number of documents retrieved by the initial query. The complexity of LSI is thus independent from the size of the document collection. The approach, on the other hand, uses only the query and the ontology to expand named entities. As a result, the complexity of our approach does not depend on to the number of documents in the collection, except for retrieving feedback documents (which depends on the retrieval model).

Comparing all of the approaches (QL, RM3 and SMERA) in our study depended on many parameters (cf. Table 4). The values of these parameters were chosen by optimizing the average performance of the measure MAP for each collection. These values were obtained after a tuning step. The experience presented in this paper corresponds to the values presented for each collection in Table 5.

4.2 Results

Table 6 presents the values obtained for the evaluation measures on the four collections and for the compared approaches. Statistically significant improvements or degradations for each couple of approaches are presented in Table 7.

In Table 7 we see that SMERA achieves statistically significant improvement in MAP compared to the use of non-expanded queries for INEX 2006, WT10g and Gov2 collections. Analyzing the test case INEX 2009 showed that 57 % of INEX 2009 queries contained at least four useful terms, larger than the average

Table 5 Selected values of the free parameters for our four test cases

	Inex 2006	Inex 2009	WT10g	Gov2
μ	**2500**	**2500**	**2500**	**2500**
n_{Smera}	20	10	30	10
n_{Rm3}	10	10	10	10
m	5	7	3	7
t	20	20	20	20
λ_{Smera}	0.8	0.8	0.5	0.8
λ_{Rm3}	0.5	0.8	0.8	0.8
k	10	5	10	5
$\alpha 1$	**0.4**	**0.4**	**0.4**	**0.4**
$\alpha 2$	**7**	**7**	**7**	**7**

Table 6 Evaluation results in MAP, P@10, MRR and ROM on the four test collections

		MAP	P@10	MRR	ROM
Inex2006	QL	33.00	53.00	81.97	83.19
	RM3	**35.96**	**55.00**	80.37	**84.61**
	SMERA	34.78	53.71	**84.81**	83.71
Inex2009	QL	34.17	**97.50**	97.79	45.89
	RM3	34.06	96.76	97.43	45.87
	SMERA	**34.41**	97.21	**98.53**	**46.18**
WT10g	QL	20.16	29.18	58.54	70.74
	RM3	20.49	29.08	56.10	71.06
	SMERA	**21.69**	**29.80**	**59.42**	**71.40**
Gov2	QL	29.41	53.51	72.36	70.57
	RM3	29.97	52.97	68.86	71.15
	SMERA	**30.82**	**56.22**	**75.84**	**71.70**

Bold values are the highest in their column

length of web queries. The MAP of the baseline (QL) obtained in this test case was the highest compared to the one obtained for INEX 2006, WT10g and GOV2. In fact, our approach is designed to improve the precision of short ambiguous queries. Expanding long queries that already have good precision has less chance to improve the performance, as it could change the order of relevant documents already retrieved by the original query. Nevertheless, SMERA obtained statistically better MAP than RM3 on this collection. This can be explained by the use of the quality filters defined in Sect. 3.1.3. Because of these filters, SMERA does not systematically expand all queries with the same number of terms; unlike RM3, which systematically adds 20 expansion terms. Most queries of the other three test cases contain from two to three useful terms (which corresponds to the general case of web queries). For these collections, SMERA had between 4.79 and 7.59 % better MAP than QL. The only

Table 7 Improvement or degradation percentage in *MAP*, *P@10*, MRR et *ROM* for each couple of approaches on the four test collections

		MAP	P@10	MRR	ROM
Inex2006	RM3/QL	+8.97*	+3.77*	−1.95	+1.71
	SMERA/QL	+5.39*	+1.40	+3.46	+0.63*
	SMERA/RM3	−3.28	−2.35	+5.52*	−1.06
Inex2009	RM3/QL	−0.32	−0.76	−0.37	+0.04
	SMERA/QL	+0.70	−0.30	+0.76	+0.63
	SMERA/RM3	+1.03	+0.47	+1.13	+0.68
WT10g	RM3/QL	+1.64	+0.34	−4.16	+0.45
	SMERA/QL	+7.59*	+2.12	+1.50	+0.93
	SMERA/RM3	+5.86*	+2.48	+5.92	+0.48
Gov2	RM3/QL	+1.90*	−1.00	−4.84	+0.82*
	SMERA/QL	+4.79*	+5.06*	+4.91*	+1.60*
	SMERA/RM3	+2.84*	+6.13*	+10.14*	+0.77

* indicates statistical significance ($p < 0,05$) for both t-test and randomization test

case in which RM3 obtained better MAP than SMERA was on INEX 2006 test case, which had the particularity of having the smallest document collection. On the other hand, this better performance in MAP of RM3 over SMERA for the case of INEX 2006 was not statistically significant.

The behavior of RM3 and SMERA in P@10 and MRR was similar to their behavior in MAP on the four test cases. Again, the expansion approaches could not obtain significant improvement in P@10 and MRR on the collection INEX 2009. But SMERA achieved significant improvement over RM3 in MRR on the collection INEX 2006, even though RM3 is better (without statistical significance) on the other measures for this collection. The positive results of SMERA on MRR for the four test cases means that it was able to find the first relevant documents in higher ranks than RM3, which is a very appreciable behavior in a web context.

Another interesting observation is the good performance of SMERA on the largest test case, Gov2. It significantly outperformed QL and RM3 in all precision measures. To better understand this observation, we explored the effect of the collection size on the behavior of RM3 and SMERA. In Fig. 6 we plotted the improvements obtained by RM3 and SMERA in P@10 over the use of non-expanded queries on the four collections.

From this figure we note the decreasing relation between the precision at rank 10 of RM3 and the collection size: the larger the collection of documents is, the less improvement RM3 achieves in P@10. Conversely, SMERA reports better improvement in precision at rank 10 with larger collections, which is also a beneficial behavior in the case of the web. The only exception for SMERA is the case of INEX 2009 because of its long queries, which is not the common case of web queries.

Even though in a web context the recall is not a priority, we think that the study of an approach's behavior from different perspectives to helps better use it in the aimed

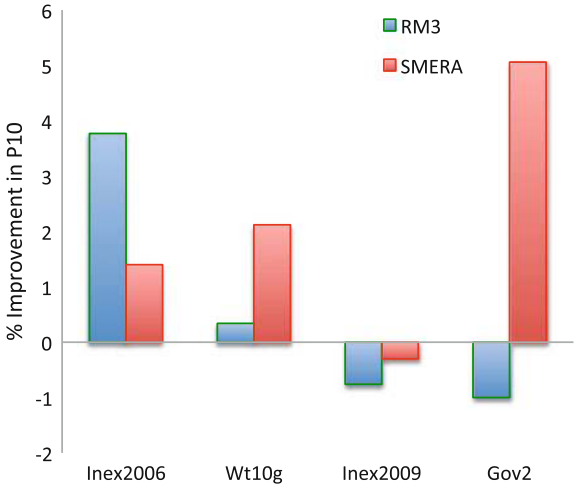

Fig. 6 Percentage of improvement in *P@10* for RM3 and SMERA on the four test collections in ascending order according to their size (in number of documents)

context. The ROM measure shows that both expansion approaches (SMERA and RM3) did not have large neither significant improvements over the baseline. This means that these approaches were not able to find more relevant documents than an approach that uses the basic non-expanded queries. This behavior is due to two main reasons: the first reason is the already high recall of the baseline on all our test cases, as can be seen in Table 8.

The second reason could be the high percentage of non judged documents among the sets of retrieved documents in our test cases (Fig. 7), which is a common but important problem with evaluation campaigns.

This means that even if expansion approaches find new relevant documents, there is a high probability that the documents found were not judged (positively or not) by an assessor.

Finally, we present the advantage of mixing two different approaches of query expansion over the use of each approach separately. While comparing SMERA to the feedback approach, we also analyzed the effect of the number of feedback documents and the number of LSI dimensions, two main parameters that are usually fixed experimentally in similar approaches. In Fig. 8, we fixed the number of feedback documents to 100 and varied the number of dimensions for the collections WT10g and INEX 2006. This performance is compared to SMERA and RM3 with the configurations mentioned in Table 5.

Table 8 The recall at 1000 for the model QL on our four test cases

	Inex2006	Inex2009	WT10g	Gov2
Recall@1000 of QL	83.85	45.95	72.03	71.05

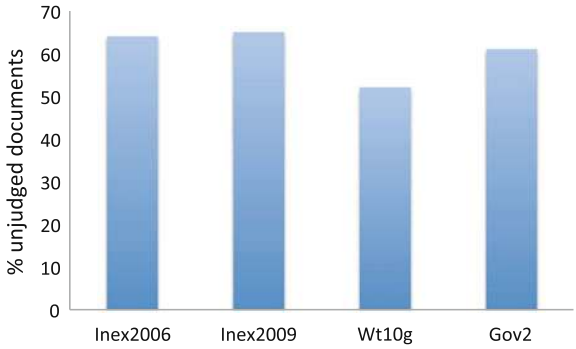

Fig. 7 The average percentage of non judged document per query for our test collections

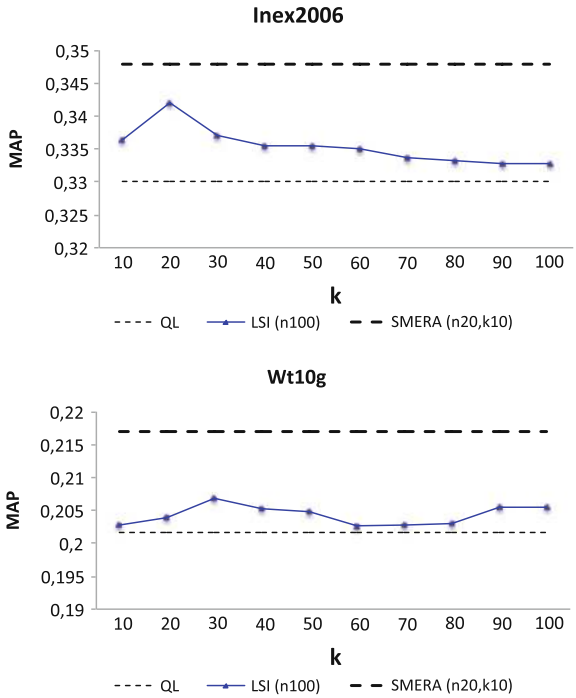

Fig. 8 Mean average precision sensibility to the number of LSI dimensions (k) for 100 feedback documents

In Fig. 8 we see that using 100 feedback documents with the feedback approach alone could enhance the recall and the precision with 30 and 20 dimensions for the WT10g and INEX 2006 collections respectively, but it was not as good as using the mixed approach of SMERA with 20 to 30 feedback documents for these two collections.

In addition to comparing SMERA to the feedback approach alone, we compared it to the use of the YAGO approach alone. For the later approach, we also considered the effect of disambiguation against the use of the most common concept corresponding to a term in the query. Fig. 9 shows that the effect of using the disambiguation or

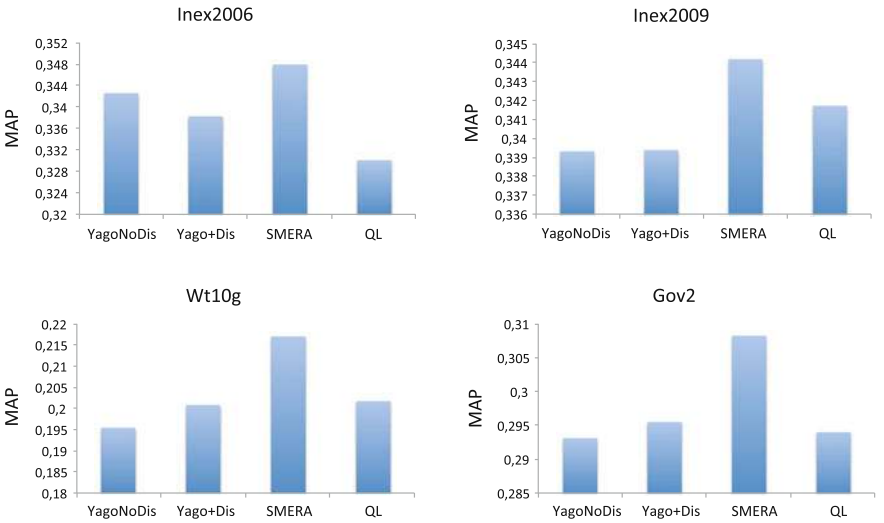

Fig. 9 Mean average precision of SMERA compared to QL, the ontology approach with disambiguation (YAGO+Dis), and the ontology approach with using the most common sense to associate concepts to query terms (YagoNoDis)

not is not stable over the collections–but it is clear that mixing the YAGO approach with LSI through SMERA has better performance in MAP than the use of the YAGO approach alone or using original queries without expansion.

5 Conclusion

In this paper we presented SMERA, a mixed approach to semantically expand and reformulate web queries. The motivation of this proposal was the lack of approaches that take into account the characteristics of web queries. More specifically, our study revealed the need of an expansion approach that considers the importance of named entities and allows an efficient, yet comprehensive, semantic representation of expanded queries. Representing concepts in a keyword query revealed the need to carefully handle the selection of expansion sets and the importance of the way in which these sets should be represented in the final query.

Evaluating our approach on four standard test collections showed the advantage of using SMERA over the use of non-expanded queries and the use of a state-of-the-art expansion method (RM3). Although not very powerful in improving the recall, our approach showed scalability and statistically significant improvements in several precision measures. The analysis of these results, and the comparison to the use of one of the proposed expansion methods in our expansion approach, suggests that SMERA is a well adapted approach for web queries' reformulation.

As a next step, we plan to investigate semantic relationships other than "rdf:label" in YAGO. The idea is to see if a sophisticated choice of the semantic relationship according to the entity type could be of interest. On the other hand, in this work we relied on the assumption that all query concepts (that we discover through our expansion approaches) have the same importance to the query. As we have seen, the approach achieved good performance even with the above assumption. We would like to explore possible solutions to weight concepts' representation, which we would obtain from resources of a different nature: a set of documents (via LSI) and an ontology (via YAGO). Finally, we are convinced of the importance of selective query expansion, which means considering the quality of added terms and not systematically expanding all query terms in the same manner. We saw this aspect investigated in information retrieval, but not explored much by query expansion approaches. Thus, testing existing quality prediction approaches and comparing them to our proposed specificity and certitude filter is an important future step to our work.

References

Audeh, B., Beaune, P., & Beigbeder, M. (2013). Recall-oriented evaluation for information retrieval systems. In: *Information Retrieval Facility Conference (IRFC), Limassol, Chypre*.

Barr, C., Jones, R., & Regelson, M. (2008). The linguistic structure of english web-search queries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1021–1030). Association for Computational Linguistics.

Bendersky, M., & Croft, W. B. (2008). Discovering key concepts in verbose queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 491–498). ACM.

Bendersky, M., Metzler, D., & Croft, W. B. (2012). Effective query formulation with multiple information sources. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (pp. 443–452). ACM.

Bendersky, M., Rey, M., & Croft, W. B. (2011). Parameterized concept weighting in verbose queries. In *SIGIR*. ACM Press.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Brandao, W., Silva, A., Moura, E., & Ziviani, N. (2011). Exploiting entity semantics for query expansion. In *IADIS International Conference WWW/Internet, Rio de Janeiro*.

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 299).

Deerwester, S., Dumais, S. T., Furnas, G. W., & Landauer, T. K. (1990). Indexing by latent semantic analysis. *Society*, *41*, 391–407.

Deveaud, R., Bonnefoy, L., & Bellot, P. (2013). Quantification et identification des concepts implicites d'une requête. In *CORIA 2013, La dixième édition de la COnférence en Recherche d'Information et Applications, Neuchâtel*.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*.

Hoffart, J., Yosef, M. A., Bordino, I., Furstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., & Weikum, G. (2011). Robust disambiguation of named entities in text. In *EMNLP 2011 Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 782–792).

Huston, S., & Croft, W. B. (2010). Evaluating verbose query processing techniques. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 291–298). ACM.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, *36*, 207–227.

Kumaran, G., & Carvalho, V. R. (2009). Reducing long queries using query quality predictors. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 564). NY, USA: ACM Press.

Lavrenko, V., & 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* (pp. 120–127). NY, USA: ACM Press.

Maxwell, K. T., & Croft, W. B. (2013). Compact query term selection using topically related text. In *Proceedings of the 36th International ACM SIGIR* (pp. 583–592).

Metzler, D., & Croft, W. B. (2004). Combining the language model and inference network approaches to retrieval. *Information Processing and Management*, *40*, 735–750.

Metzler, D., & Croft, W. B. (2005). A Markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 472). NY, USA: ACM Press.

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

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 275–281). ACM.

Qiu, Y., & Frei, H. (1993). Concept based query expansion. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Informaion Retrieval* (Vol. 11, p. 212). NY: ACM.

Rocchio, J. J., & Salton, G. (1965). Information search optimization and iterative retrieval techniques. In *Fall Joint Computer Conference* (pp. 293–305).

Shah, C., & Croft, W. B. (2004). Evaluating high accuracy retrieval techniques chirag shah. In *SIGIR*. ACM Press.

Strohman, T., Metzler, D., Turtle, H., & Croft, W. (2004). Indri: A language-model based search engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis*.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 697–706). ACM.

Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In *SIGIR 1994*. ACM Press.

Xu, Y., Ding, F., & Wang, B. (2008). Entity-based query reformulation using wikipedia. In *Proceeding of the 17th ACM Conference on Information and Knowledge Mining - CIKM 2008* (p. 1441). NY, USA: ACM Press.

Zhao, L., & Callan, J. (2010). Term necessity prediction. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 259–268). ACM.

Zobel, J. (2004). Questioning query expansion: An examination of behaviour and parameters. In *SIGIR*. ACM Press.