Exploiting Wikipedia Structure for Short Query Expansion in Cultural Heritage

Mohannad ALMASRI — Jean-Pierre CHEVALLET — Catherine BERRUT

Grenoble Alpes University, LIG laboratory, MRIM group
{*Mohannad.almasri,jean-pierre.chevallet,catherine.berrut*}*@imag.fr*

ABSTRACT. *This paper deals with the short and precise queries problem. Short and precise queries do not have sufficient information to be non ambiguous. Pseudo-relevance feedback (PRF) is an effective technique to improve retrieval performance by expanding a user query. However, this collection based expansion method does not work well in the case of short queries. Therefore, we present instead of PRF, a semantic query expansion method based on Wikipedia as external knowledge. We expand short queries by semantically related terms extracted from Wikipedia. We propose and study the effectiveness of three variations for expansion terms selection. We incorporate the expansion terms into the original query and adapt language models to evaluate the expanded queries. Experiments on CLEF cultural heritage corpora show significant improvement in the retrieval performance. We show that the number of expansion terms has an important impact on the precision improvement.*

RÉSUMÉ. *Cet article aborde le problème des requêtes courtes et précises, qui n'ont pas suffisamment d'informations pour être non ambiguës. Le pseudo-relevance feedback (PRF) est une technique efficace pour améliorer la performance de ces requêtes courtes par l'ajout de termes à la requête. Cependant, cette méthode ne récupère que les termes des documents les plus pertinents de la collection. Si les réponses initiales ne sont pas correctes, comme c'est le cas pour des requêtes courtes, cette expansion ne fonctionnera pas. Par conséquent, nous présentons à la place du PRF, une méthode d'expansion sémantique des requêtes basée sur Wikipedia. Nous étendons requêtes courtes par des termes sémantiquement liés. Nous adaptons les modèles de langue pour évaluer les requêtes étendues. Les expérimentations sur une corpus CLEF du patrimoine culturel montrent une amélioration significative de la performance. Nous montrons que le nombre de termes d'expansion a un impact important sur l'amélioration de la précision.*

KEYWORDS: *Query Expansion; External Resources; Language Model; Cultural Heritage*

MOTS-CLÉS : *Extension des Requêtes ; Ressources Externes ; Modèle de Langage*

CORIA 2014, pp. 287–302, Nancy, 18-21 mars 2014

1. Introduction

Short and precise queries have no sufficient information to be non ambiguous. For example, in the cultural heritage domain, the query "*last supper*". A classical information retrieval (IR) model will retrieve documents containing these two words or one of them without giving any attention to the particular meaning of this query in the Christian religion. This information is difficult to infer from the query only. However, adding some semantically related terms [1] to this query, like "*jesus*", "*crucifixion*", "*twelve apostles*", "*judas*" etc., could clarify the meaning of this query and enhance the ability of IR models to retrieve the relevant documents.

Another example from the same domain, the query "*silent film*" which searches for documents on history of silent film, actors and directors. A document talking about "*charlie chaplin*", for instance, is a relevant document to this query. However, a classical IR model is incapable to retrieve this document without an additional information about this link between: "*silent film*" and "*charlie chaplin*".

Pseudo-Relevance Feedback (PRF) is a method for query expansion using terms extracted from top k retrieved documents. However, if the top retrieved documents for a given query contain a few number of relevant documents, which is the case of short queries, then selected terms, using PRF, will not be strongly related to the original query. As a result, retrieval performance for the expanded query is not better than the original query (Xu *et al.*, 2009; Li *et al.*, 2007; Diaz and Metzler, 2006; Akasereh *et al.*, 2012; Akasereh *et al.*, 2013).

In this study, we use corpora from cultural heritage field. Cultural heritage is one of the most valuable resources that store the accumulated knowledge of humankind. Nowadays, many organizations, such as museums and libraries, own huge collections providing historical cultural data. Seekers querying these information, normally use short queries that include named entities, e.g. person, place, event, etc. Therefore, we present here a semantic query expansion method in order to overcome the short query problem. Our method proposes to select, for a given query, some semantically related terms from an external resource different from corpus. We think that Wikipedia is a convenient knowledge source, because it is a large knowledge containing a huge number of articles about named entities. Our interest is to exploit the content of Wikipedia and its internal structure in order to expand short queries. Then, incorporate the expansion terms into the original query and adapt language models to evaluate expanded queries. We claim that this semantic expansion improves the retrieval performance for short queries.

The rest of paper is organized as follows: section 2 describes some existing approaches in query expansion; section 3 presents our semantic query expansion method; our experimental set-up and the empirical results are presented in section 4; finally, section 5 concludes the paper.

1. term: one or more words.

2. Related Works

Query expansion has been widely studied as an efficient way to resolve the short query problem (Cui *et al.*, 2002). PRF approach considers all top retrieved document and their terms as the expansion candidates. However, in the short and precise query (ex: named entity) the top answer list contain too many non-relevant documents. Besides, documents in the feedback set although containing relevant information, are sometimes partially related to the topic, and therefore yield bad expansion terms (Macdonald and Ounis, 2007). Later works try to detect good expansion terms using a trained classifier (Cao *et al.*, 2008) or select relevant documents for feedback by an active learning algorithm (Xu and Akella, 2008). While these previous works use the collection itself for pseudo feedback, the feedback can come from different external resources. Participants in TREC robust retrieval track have successfully used large web search engine results for pseudo feedback (Voorhees, 2005).

Query logs is an other example that exploit external resources in query expansion (Billerbeck *et al.*, 2003). Beeferman et al. (Beeferman and Berger, 2000) use query log to build a bipartite graph where its vertices queries from log and clicked URL. Queries and URLs are connected by edges (clicks). Then, agglomerative clustering is used to identify related queries and URLs. Wen et al. (JI-RONG WEN, 2002) use clicked documents to compute the similarities between queries. Sahami et al. (Sahami and Heilman, 2006) propose a web-based method for measuring the similarity between queries by building a context vector to each query from the top search engine results. In our paper, we use a graph-based similarity between Wikipedia articles. We therefore use this similarity to select expansion terms from Wikipedia as we present it as a directed graph of articles.

Bendersky et al. (Bendersky *et al.*, 2012) propose a framework for query expansion that use different external information sources like web collections and Wikipedia. Diaz zt al. (Diaz and Metzler, 2006) study the use of several web collections as an external resource for enhancing the estimation of relevance model. Li et al. (Li *et al.*, 2007) use Wikipedia as an external corpus to expand short query. Similarly, Xu et al. (Xu *et al.*, 2009) propose an approach for pseudo relevance feedback based on Wikipedia as an external resource. Our work belongs to this category and we use Wikipedia as an external resource for query expansion. However, we concentrate on using Wikipedia structure in our proposal instead of consider it as a collection of articles without taking into account its structure or links between its articles.

Collins-Thompson et al. (Collins-Thompson and Callan, 2005) exploit different information sources: Wordnet, Krovetz stemmer, general word association and retrieved documents to build the term lexical and semantic relationship graphs, used to calculate the most related terms to expand original query. However, this linguistic based solution is not suitable for cultural heritage domain which deals with named entities.

3. Semantic Query Expansion

The key point in any query expansion method is to generate expansion terms that can improve retrieval performance. In PRF top k documents are considered as the expansion terms source. In our semantic query expansion method, expansion terms source is defined over Wikipedia. Therefore, we first present our representation of Wikipedia. Then, we precisely define our expansion terms source and the criteria to select expansion terms from this source based on our Wikipedia representation. After that, we explain the different steps of our semantic query expansion method. Last, we explain the incorporation of our expansion terms into language models.

3.1. *Wikipedia as a Graph*

Wikipedia is an encyclopedia that represents a very large, high quality, and valuable knowledge source in natural language. Moreover, Wikipedia is also a hypertext in which each Wikipedia article can refer to other Wikipedia article using hyperlinks. We consider only *internal links*, which are links that target an other Wikipedia article.

We represent Wikipedia articles as a directed graph $G(A, L)$ of articles A connected by links $L \subseteq A \times A$. Each article $a \in A$ is a description of an object, an entity, an historical fact, etc. (entitled as $title(a) \in T$), where T is a set of terms. We consider that every article title is a term, eventually composed of one or more words. For instance: "*Painting*", "*Last Supper*", "*Santa Maria delle Grazie (Milan)*", "*New Testament places associated with Jesus*".

Furthermore, each article contains links to other articles. L is relations between articles defined on $A \times A$ where (a_1, a_2) means the article a_1 have a link to the article a_2. In the rest of this paper, we use:

$$\begin{array}{rcl} I, O & : & A \rightarrow 2^A \\ I(a) & = & \{x \in A | (x, a) \in L\} \\ O(a) & = & \{x \in A | (a, x) \in L\} \end{array}$$

where $I(a)$ is the set of articles that point to a (Incoming Links) , and $O(a)$ is the set of articles that a points to (Outgoing Links).

Moreover, we propose to weight the Wikipedia articles graph, knowing that each article describes a particular object, entity or notion, and each link proposes a navigation to a semantically related article. For this reason, we propose to evaluate the strength of all these links and define a semantic similarity between two articles. For that, we consider that two Wikipedia articles are semantically similar, if they share a similar link context, i.e. if they have similar incoming and outgoing link sets.

Hence, two articles a_1 and a_2 in A are semantically similar if they share articles that point to them, and if they share articles that a_1 and a_2 point to. Then, we propose the following semantic similarity:

$$SIM(a1, a2) = \frac{|I(a_1) \cap I(a_2)| + |O(a_1) \cap O(a_2)|}{|I(a_1) \cup O(a_1)| + |I(a_2) \cup O(a_2)|} \quad [1]$$

where $I(a)$ are incoming links to article a, and $O(a)$ are outgoing links from a.

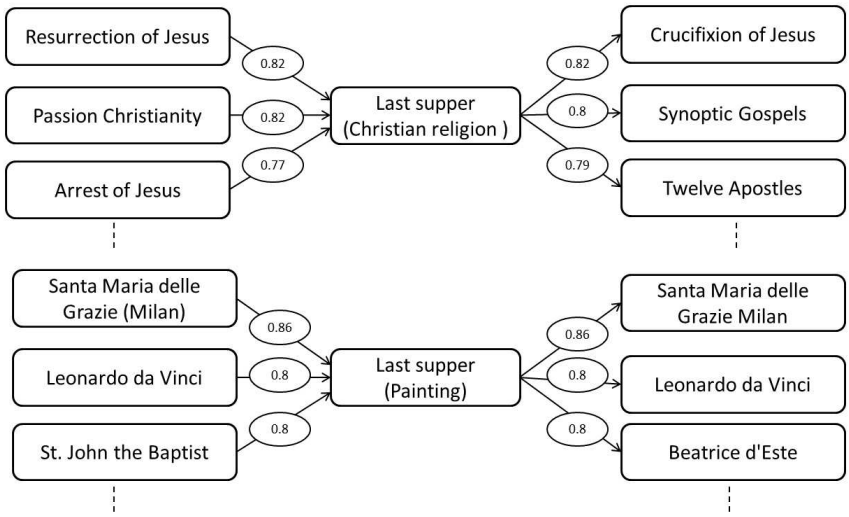


Figure 1: Example of two articles in Wikipedia with the same title "*Last Supper*". Each article corresponds a sense of the term "*Last Supper*" in Wikipedia. Each article is connected with two set of articles:incoming links I and outgoing links O. Each ellipse between two linked articles corresponds the semantic similarity between these two articles calculated by Eq.[1].

3.2. *Wikipedia Utilization*

In this section, we explain the exploiting of Wikipedia graph in order to find a set of semantically related terms to a given term. In other words, we now define our expansion term source which we use in our semantic query expansion method. Therefore, we precisely present the notion of term and how we collect a set of related terms based on Wikipedia graph.

3.2.1. *Term Polysemy*

In Wikipedia two different articles may share the same title. In that case, each article refers to one special single sense of this term. For example, a term "*Last supper*"

corresponds several articles entitled by "*Last supper*" in Wikipedia. Figure 1 shows two of them. In one article (or sense), last supper is the final meal according to Christian belief, the other article describes last supper as the painting of Leonardo da Vinci. We define $S(t)$ the set of senses of a term $t \in T$, or in other words, the set of articles entitled by t:

$$S(t) = \{a \in A | title(a) = t\} \quad [2]$$

Each sense of t is an article a that can be the target of a hyperlink. Hence, we can calculate a notion of "*popularity*" of a, and consequently of a sense, just by computing the probability to choose a particular sense a, for a given term t among the other sense from $S(t)$. We estimate the probability of any term t that is the anchor of an internal link, to be linked to an article a by the maximum likelihood:

$$P(a|t) = \frac{\mid I(a) \mid}{\sum_{x \in S(t)} \mid I(x) \mid} \quad : \quad a \in S(t) \quad [3]$$

After identifying the term t, we now search the most n similar terms to t from Wikipedia, where the main criteria for the number of selected terms for each sense $a \in S(t)$ is its probability $P(a|t)$. The more probable the sense is the more it contributes in term selection. Terms for each sense $a \in S(t)$ comes from the titles of linked articles to a. Article title precisely identifies the subject, and is short, natural, and recognizable (Wikipedia, 2013). Formally, we define the function $topSimTerm(t, n)$:

$$topSimTerm \quad : \quad T \times \mathbb{N} \rightarrow 2^{T \times \mathbb{R}}$$

$$topSimTerm(t, n) = \bigcup_{a \in S(t)} topSimTitle(a, \lceil P(a|t) \times n \rceil) \quad [4]$$

where the function $topSimTitle$ returns the most similar article titles to a as defined in the following section.

3.2.2. *Article linkage*

As each article in Wikipedia is linked to some other articles, we define a function $linked(a)$ that returns the set of linked articles with their similarities to $a \in A$.

$$linked \quad : \quad A \rightarrow 2^{A \times \mathbb{R}}$$

where $linked(a)$, for instance, selects articles from $I(a)$, $O(a)$, $I(a) \cup O(a)$, or any other possibility based on the Wikipedia graph. Suppose the example in Figure 1, where $linked(a) = I(a)$, then $linked$(Last supper(Christian religion)) $=$ $\{$(Resurrection of Jesus, 0.82), (Passion Christianity, 0.82) , (Arrest of Jesus, 0.77)$\}$. Each pair $(x, SIM(a, x))$ from $linked(a)$ contains a similarity value calculated using Eq.[1]. We define the function $topSim(a, n)$, which returns the most n similar articles to a from $linked(a)$ based on their similarity.

$$topSim \quad : \quad A \times \mathbb{N} \rightarrow 2^{A \times \mathbb{R}}$$

Finally, based on $topSim$, we define $topSimTitle$ which returns the most n similar article titles to a from $linked(a)$ in order to use it in the Eq.[4]. Each term takes the same similarity of its attached article.

$$\begin{array}{rcl} topSimTitle & : & A \times \mathbb{N} \rightarrow 2^{T \times \mathbb{R}} \\ topSimTitle(a,n) & = & \{(title(x), sim) | (x, sim) \in topSim(a,n)\} \end{array}$$

Return to same example in Figure 1, where $linked(a) = I(a)$ we obtain: $topSimTitle$(Last supper(Christian religion), 2) $= \{$(“*Resurrection of Jesus*”, 0.82) , (“*Passion Christianity*”, 0.82)$\}$.

We consider, in this section, Wikipedia article titles as a source of expansion terms. Article title precisely identifies the subject, and is short, natural, and recognizable. Article titles are rich source of information for query expansion. In addition, titles appear as a highlighted text within other Wikipedia articles. This appearance of titles in other articles carries an important semantic relation as Wikipedia articles are manually created. We explain in this section the process to obtain a set of semantically related terms to a given term. We move in the next section to our query expansion method. We therefore define a query by means of terms (or article titles) and then use the previous process in order to find related terms to a given query for expanding it.

3.3. *Building Expanded Query*

We now present our method for aggregating related articles to a given query over Wikipedia graph in order to expand this query. We define two types of queries according to their relation to Wikipedia: 1) simple queries about one named entity or contain one term, 2) compound queries about multiple named entities or contain multiple terms.

3.3.1. *Simple Query vs Compound Query*

In the context of text retrieval, a user formulates her need by a query q containing a sequence of words “$w_1 w_2 ... w_{|q|}$”. For example, the sequence “*last supper*” is composed of two words, and it is different to the sequence “*supper last*”. Based on this representation, we formally define the mapping of a sequence of words q into Wikipedia. Given a query $q =$ “$w_1, w_2, ..., w_{|q|}$”:

– We denote by $w_{x \rightarrow y}$ a consecutive sub sequence of words of q, for $x, y \in [1; |q|]$.

– We define a function $M(q)$ that maps a query q into the largest term as follows:

$$\begin{array}{rcl} M(q) & = & \{w_{x \rightarrow y} | \exists a \in A : title(a) = w_{x \rightarrow y} \\ & & \wedge \nexists w_{x' \rightarrow y'} > w_{x \rightarrow y} : title(a) = w_{x' \rightarrow y'}\} \end{array}$$

This function M maps a query q into a set of Wikipedia article titles.

– Based on M, which maps q into a collection of Wikipedia titles, we define:

- Simple query q_s about one named entity which satisfies $M(q_s) = \{w_{1 \rightarrow |q_s|}\}$. For instance, the query “*silent film*” is an example of a simple query.

- Compound query q_c about multiple named entities which satisfies $|M(q_c)| = k > 1$. For example, the query q =“*hiroshima and nagasaki*” contains two named entities: first entity is “*hiroshima*”, where second entity is: “*nagasaki*”.

3.3.2. *Expand Simple Query*

The input is a simple query q_s, and the number of expansion terms n. The output is a weighted set of n related terms added to q_s to obtain the expanded query q_{exp}.

$$q_{exp} = q_s \cup q'_s \text{ where } q'_s \text{ has } n \text{ terms}$$

Given a query q_s:

– Collect all Wikipedia articles $S(q_s)$ entitled by q_s.

– The expansion terms for a query q_s are the union of article titles comes from linked articles to each $a \in S(q_s)$, using the function $topSimTerm$, see Eq.[4]. Thus, we define the expanded query q_{exp} as follows:

$$q_{exp} = q_s \cup q'_s$$
$$q'_s = \{t|(t, sim) \in topSimTerm(q_s, n)\}$$

– Terms are weighted by a value between $[0, 1]$, which reflects the importance of each term. For each $t \in q_{exp}$:

$$weight(t, q_{exp}) = \begin{cases} 1 & \text{if } t \in q_s \\ \alpha \times sim & \text{if } (t, sim) \in topSimTerm(q_s, n) \end{cases} \quad [5]$$

Where $\alpha \in [0, 1]$ is a tuning parameter determines the importance of expansion terms.

3.3.3. *Expand Compound Query*

We now present our method in order to capture compound queries containing multiple entities. Formally, assume a compound query q_c with k entities or terms.

$$\text{we denote by: } q_i = w_{x \to y}, \forall w_{x \to y} \in M(q_c)$$

The input of our method is q_c, and the number of expansion terms n. The output is a weighted set of n related terms added to q_c to obtain the expanded query q_{exp}.

$$q_{exp} = q_c \cup q'_c \text{ where } q'_c \text{ has } n \text{ terms}$$

Given a query q_c:

– For each entity q_i collect all Wikipedia articles $S(q_i)$ entitled by q_i.

– The expansion terms for each entity q_i are the union of article titles comes from linked articles to each $a \in S(q_i)$. Knowing that each entity q_i contributes equally by n/k terms in the expansion of q_c. Thus, we define the expanded query q_{exp} for a compound query q_c as follows:

$$q_{exp} = q_c \cup q'_c$$
$$q'_c = \bigcup_{i=1}^{k} q'_i$$
$$q'_i = \{t|(t, sim) \in topSimTerm(q_i, n/k)\}$$

– Terms are weighted in the expanded query q_{exp}. In this case, expansion term weight depends on the entity q_i where a term comes from. For each $t \in q_{exp}$:

$$weight(t, q_{exp}) = \begin{cases} 1 & \text{if } t \in q_s \\ \alpha \times sim & \text{if } (t, sim) \in topSimTerm(q_i, n/k) \end{cases} \quad [6]$$

Where $\alpha \in [0, 1]$ is a tuning parameter determines the importance of expansion terms.

3.4. *Retrieval*

As we mentioned in the previous section, we choose the highest similar articles for a given query q, and we use their titles as expansion terms for this query. Now, we move from terms space into unigram or word space as we plan to modify unigram language model in order to incorporate expansion terms into language models. As a result, we lexicalize these titles in order to get their words. Every word within a term takes the same weight of this term. For example, assume the query $q =$"*last supper*", and "*twelve apostles*" is a term for expanding this query, with the semantic similarity: 0.79. Therefore, word weights in the expanded query q_{exp} for the previous example are: $\{weight(last, q_{exp}) = 1$, $weight(supper, q_{exp}) = 1$, $weight(twelve, q_{exp}) = \alpha \times 0.79$, $weight(apostles, q_{exp}) = \alpha \times 0.79\}$.

Our retrieval model runs queries which contain the original terms as well as the expansion terms. Therefore, we propose to modify language model to take into account the difference between original and expansion terms. The reason behind this modification is to promote original terms above expansion terms in the expanded query by capturing the semantic relation between the original query terms and expansion terms. That leads us to use a similar idea from statistical translation language model (Karimzadehgan and Zhai, 2010), where we consider each expansion term as a semantic probable translation of an original query term. The basic idea of language model is to assume that a query q, which is generated by a probabilistic model based on a document d. Jelinek-Mercer and Dirichlet are two variation of language models (Zhai and Lafferty, 2004). Therefore, we replace maximum likelihood $P_{ml}(w|d)$ in these two, by a new probability that consider the semantic distance between original and expansion query terms.

– Jelinek-Mercer language model is defined by the following formula:

$$RSV(q, d) = |q| \times ln(\lambda) + \sum_{w \in d \cap q} tf_{w,q} \times ln((1 - \lambda)p_{ml}(w|d) + \lambda p(w|C))$$

– Dirichlet language model is defined by the following formula::

$$RSV(q, d) = |q| \times ln(\tfrac{\mu}{|d|+\mu}) + \sum_{w \in d \cap q} tf_{w,q} \times ln(\tfrac{|d|}{|d|+\mu} p_{ml}(w|d) + \tfrac{\mu}{|d|+\mu} p(w|C))$$

– We replace maximum likelihood $p_{ml}(w|d)$ in the previous two equations in our expanded query q_{exp} by the probability $P_{exp}(w|d)$ defined by:

$$P_{exp}(w|d) = p(w|q_{exp}) \times P_{ml}(w|d) \quad [7]$$

where $p(w|q_{exp})$ the probability of translation for a word w in the expanded query q_{exp}. This probability depends on the semantic distance or similarity between the term t that w belongs to and the original query q. Thus, we estimate this probability as follows:

$$\forall w \in t, t \in q_{exp} : \quad p(w|q_{exp}) = weight(t, q_{exp})$$

As we see, the probability of w depends on the weight of the term t that w belongs to, see Eq.5 and Eq.6. If t belongs to the original query then this probability equals 1 and we return to the normal language model. However, if w belongs to one of expansion terms then this probability depends on the semantic distance of the term containing this word and the original query.

4. Experiments

4.1. *Target Collections*

Experiments are conducted using two CLEF collections for cultural heritage: CHIC2012 and CHIC2013 English collections. Each collection contains 1,107,176 short documents [2], and 25 topics about named entities. We use only topic titles in our evaluation. These two collections correspond semantic query enrichment task . This task differs to ad hoc retrieval task which contains 50 queries in each collection. Average document length in these two collections is 54 words. Documents are retrieved using two smoothing methods of language models: Jelinek-Mercer(JM) and Dirichlet(DIR). We use to achieve our experiments Indri, an open source search engine (Strohman *et al.*, 2004).

Table 1: Baselines using Jelinek-Mercer (JM) and Dirichlet (DIR) models, with language model (LM) and relevance language model (RLM).

	CHIC2012		CHIC2013	
	MAP		MAP	
Method	JM	DIR	JM	DIR
LM	0.3708	0.3768	0.3552	0.3627
RLM	0.3688	0.3724	0.3549	0.3621

2. We have the same documents in both CHIC2012 and CHIC2013.

Table 2: Semantic query expansion (SQE) results using incoming links I, outgoing links O, or Both IO, using Jelinek-Mercer. † indicates significant improvement over LM and RLM using paired t-test with $p < 0.05$. The percentage is for the difference between our expansion results and LM as it is the best baseline. Bold values shows best MAP between different values of number of expansion terms n.

JM	Links	n	CHIC2012		CHIC2013	
			MAP	Gain	MAP	Gain
LM	-	0	0.3708	-	0.3552	-
RLM	0	-	0.3688	-	0.3549	-
SQE	I	5	0.4231	+14%	0.4100	+15%
		10	**0.4262**†	+15%	**0.4199**†	+18%
		15	0.4177	+13%	0.4073	+15%
		20	0.4175	+13%	0.3983	+12%
		25	0.4089	+10%	0.3926	+11%
		30	0.4037	+9%	0.3878	+9%
	O	5	0.4307†	+16%	0.4099	+15%
		10	**0.4317**†	+16%	**0.4213**†	+19%
		15	0.4225	+14%	0.4068	+15%
		20	0.4207	+13%	0.4003	+13%
		25	0.4135	+12%	0.3957	+11%
		30	0.4094	+10%	0.3893	+10%
	I+O	5	0.4210	+14%	0.4099	+8%
		10	0.4307†	+16%	0.4127	+16%
		15	0.4311†	+16%	0.4158	+17%
		20	**0.4355**†	+17%	**0.4185**†	+18%
		25	0.4235	+14%	0.4123	+16%
		30	0.4162	+12%	0.4073	+15%

4.2. *External Knowledge*

We consider Wikipedia as an external knowledge for expanding our queries. We selected Wikipedia because it is a large knowledge containing a huge number of articles about named entities. Wikipedia covers 90% of our topics, i.e. 90% of topics in these two collections correspond at least one Wikipedia article. We use Wikipedia-Miner API [3] in order to exploit Wikipedia's knowledge in our expansion.

3. Wikipedia-Miner is a toolkit for tapping the rich semantics encoded within Wikipedia http://wikipedia-miner.cms.waikato.ac.nz/

4.3. *Training and Baseline*

To evaluate the different variants of our expansion method, 2-fold cross-validation are performed by partitioning the topics into two sets. First set, contains topics from CHIC2012, and second set, contains topics from CHIC2013. Then, testing phase to CHIC2012 collection use the optimal parameters tuned from CHIC2013, and vice versa. In order to find the best parameters setting we sweep over values for the number of expansion terms $n \in \{5, 10, 15, 20, 25, 30\}$, the tuning parameter $\alpha \in \{0.1, ..., 1.0\}$ used in Eq.5 and Eq.6. We also sweep over three possibilities for the function $linked$ introduced in section 3.2.2.

$$\begin{array}{lcll} linked(a) & = & I(a) & \text{OR} \\ linked(a) & = & O(a) & \text{OR} \\ linked(a) & = & I(a) \cup O(a) & \end{array}$$

We optimize our method using mean average precision MAP as a target metric. The baselines of our experiments are Dirichlet and Jelinek-Mercer language models (LM) (Zhai and Lafferty, 2004) and relevance language model (RLM) (Lavrenko and Croft, 2001). In case of RLM, we sweep over the number of terms $\{5, 10, 15, 20, 25, 30\}$ and the number of documents $\{5, 10, 15, 20, 25, 30\}$. Besides, we also consider best result of the evaluation campaign. Best MAP in CHIC2012 for the corresponding task is 0.34. Table 1 shows our baselines for the two models used in our experiments: Jelinek-Mercer and Dirichlet. We can see from this table that using relevance models does not help to enhance the retrieval performance because of short queries and documents in the two target collections: CHIC2012 and CHIC2013. These results confirm the experiments made by (Akasereh *et al.*, 2012; Akasereh *et al.*, 2013) using several PRF retrieval settings, on these two collections.

4.4. *Results*

We use, in our evaluation, MAP, as mentioned before. We test three possibilities for the function $linked$ in semantic query expansion (SQE) using $\{I, O, IO\}$. Using I means that all of n expansion terms come from incoming links I, using O means that all of n expansion terms come from outgoing links O, while using IO means that n expansion terms come from both incoming links and outgoing links. Results in Table 2 and Table 3 are obtained with best value of the tuning parameter $\alpha = 0.3$.

Results of our three variation $\{I, O, IO\}$ using Jelinek-Mercer are summarized in Table 2. Results for Dirichlet are summarized in Table 3. We see in these tables the change in the number of expansion terms $n \in \{5, 10, 15, 20, 25, 30\}$ versus the change in Mean Average Precision for our three variations of expansions and the two target collections. We first observe the consistent performance improvement achieved which confirms our belief that using Wikipedia structure for query expansion improves relevance model estimation. Second, the improvement is correlated with the variation used for selecting expansion terms and the number of expansion terms. We distinct two cases:

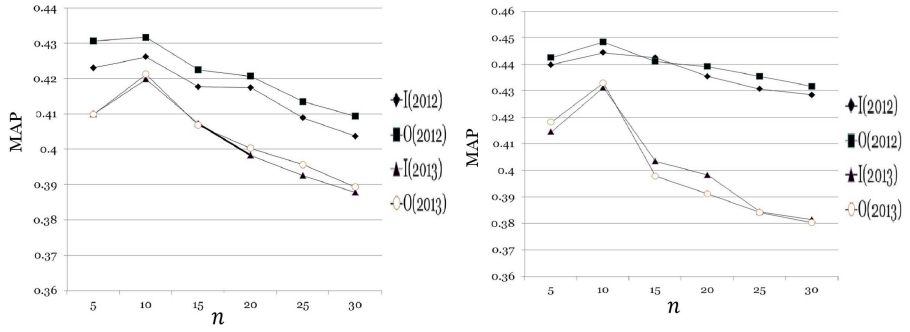
Table 3: Semantic query expansion (SQE) results using incoming links I, outgoing links O, or Both IO, using Dirichlet. † indicates significant improvement over LM and RLM using paired t-test with $p < 0.05$. The percentage is for the difference between our expansion results and LM as it is the best baseline. Bold values shows best MAP between different values of number of expansion terms n.

DIR	Links	n	CHIC2012		CHIC2013	
			MAP	Gain	MAP	Gain
LM	-	0	0.3768	-	0.3627	-
RLM	0	-	0.3688	-	0.3621	-
SQE	I	5	0.4399	17%	0.4146	14%
		10	**0.4445**†	18%	**0.4313**†	19%
		15	0.4426†	17%	0.4035	11%
		20	0.4355	16%	0.3983	10%
		25	0.4308	14%	0.3845	6%
		30	0.4285	14%	0.3815	5%
	O	5	0.4426†	17%	0.4182	15%
		10	**0.4485**†	19%	**0.4330**†	19%
		15	0.4411†	17%	0.3979	10%
		20	0.4392	17%	0.3912	8%
		25	0.4355	16%	0.3842	6%
		30	0.4317	15%	0.3803	5%
	I+O	5	0.4322	15%	0.4148	14%
		10	0.4375	16%	0.4222	16%
		15	0.4447†	18%	0.4257	17%
		20	**0.4481**†	19%	**0.4309**†	19%
		25	0.4324	15%	0.4161	15%
		30	0.4304	14%	0.4027	11%

– Expansion using only *incoming links* I or *outgoing links* O: these two variations behave similarly with the change of expansion terms number. We see a slight difference in performance between them. In addition, the best MAP improvement, using different number of expansion terms, is achieved at 10 terms. After 10, MAP improvement start to decrease systematically due to the increasing of noise generated by using a bigger number of expansion terms.

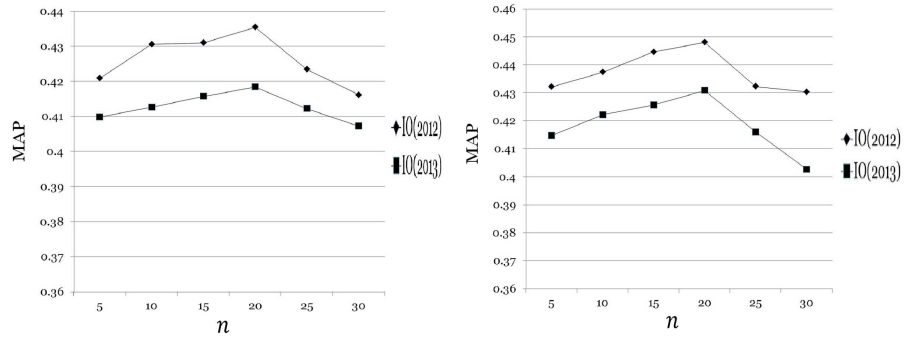
– Expansion using both *incoming* and *outgoing links* IO: we observe that the best MAP obtained using 20 expansion terms(10 from I and 10 from O). In this case, MAP improvement start to decrease when we use more than 20 expansion terms.

Results reported in the previous tables are depicted into Figure 2 and Figure 3. We see in these two figures retrieval performance in MAP as a function of number of expansion terms n. Figures 2a and 2b shows MAP changes using Jelinek-Mercer and Dirichlet,respectively, for the two variation: expansion terms from incoming links I or from outgoing links O, and for the two target collecton: CHIC2012 and CHIC2013.



(a) Jelinek-Mercer (b) Dirichlet

Figure 2: MAP as a function of number of expansion terms n using Jelinek-Mercer and Dirichlet retrieval models for the two collections: CHIC2012 and CHIC2013 and the expansion terms from incoming links I or outgoing links O.



(a) Jelinek-Mercer (b) Dirichlet

Figure 3: MAP as a function of number of expansion terms n using Jelinek-Mercer and Dirichlet retrieval models for the two collections: CHIC2012 and CHIC2013 and the expansion terms from both incoming links and outgoing links IO.

We observe that these two variations have a similar behavior over the two collections and the two retrieval models. Figures 3a and 3b shows MAP changes using Jelinek-Mercer and Dirichlet, respectively, for our third variation: expansion terms from both incoming links and outgoing links IO, and for the two target collectons: CHIC2012 and CHIC2013.

Finally, semantic query expansion (SQE) using Wikipedia structure is statistically significant better than query likelihood and relevance language model. We have a slight difference in MAP between the three possibilities of selecting expansion terms:

incoming links, outgoing links, or both. Dirichlet smoothing gives better performance in MAP than Jelinek-Mercer smoothing over all our experiments. Table 4 shows our best setting for our method.

Table 4: Best parameters setting to our semantic query expansion (SQE) over the two target collections CHIC2012 and CHIC2013: model, expansion links, α, and n.

Collection	Retrieval Model	Expansion Links	α	n	MAP
CHIC2012	Dirichlet	outgoing links O	0.3	10	0.4485
CHIC2013	Dirichlet	outgoing links O	0.3	10	0.4330

5. Conclusions

In this paper, we explore the use of Wikipedia for semantic query expansion. We propose three variants for selecting terms from Wikipedia. These variants are completely based on Wikipedia structure. We evaluate these variants on two collection CHIC2012 and CHIC2013. Our experiments results show that our method carry out a significant improvement on retrieval performance. We use for now, only titles of articles and our future investigation is to study the impact of using article abstract and full-text. Moreover, we plan to combine additional similarity metrics rather than Eq.1. Besides, we are going to merge between internal and external evidences for query expansion, because of the distribution of senses in the target collection is different to their distribution in the external resource.

6. References

Akasereh M., Naji N., Savoy J., "UniNE at CLEF 2012.", *in* P. Forner, J. Karlgren, C. Womser-Hacker (eds), *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

Akasereh M., Naji N., Savoy J., "UniNE at CLEF 2013.", *CLEF (Online Working Notes/Labs/Workshop)*, 2013.

Beeferman D., Berger A., "Agglomerative Clustering of a Search Engine Query Log", *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, ACM, New York, NY, USA, p. 407-416, 2000.

Bendersky M., Metzler D., Croft W. B., "Effective Query Formulation with Multiple Information Sources", *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, ACM, New York, NY, USA, p. 443-452, 2012.

Billerbeck B., Scholer F., Williams H. E., Zobel J., "Query Expansion Using Associated Queries", *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, ACM, New York, NY, USA, p. 2-9, 2003.

Cao G., Nie J.-Y., Gao J., Robertson S., "Selecting Good Expansion Terms for Pseudo-relevance Feedback", *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, ACM, New York, NY, USA, p. 243-250, 2008.

Collins-Thompson K., Callan J., "Query Expansion Using Random Walk Models", *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, ACM, New York, NY, USA, p. 704-711, 2005.

Cui H., Wen J.-R., Nie J.-Y., Ma W.-Y., "Probabilistic Query Expansion Using Query Logs", *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, ACM, New York, NY, USA, p. 325-332, 2002.

Diaz F., Metzler D., "Improving the Estimation of Relevance Models Using Large External Corpora", *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, ACM, New York, NY, USA, p. 154-161, 2006.

JI-RONG WEN JIAN-YUN NIE H.-J. Z., "Query clustering using user logs", *ACM Trans. Inf. Syst.*, vol. 20, n° 1, p. 59-81, 2002.

Karimzadehgan M., Zhai C., "Estimation of Statistical Translation Models Based on Mutual Information for Ad Hoc Information Retrieval", *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, ACM, New York, NY, USA, p. 323-330, 2010.

Lavrenko V., Croft W. B., "Relevance Based Language Models", *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, ACM, New York, NY, USA, p. 120-127, 2001.

Li Y., Luk W. P. R., Ho K. S. E., Chung F. L. K., "Improving Weak Ad-hoc Queries Using Wikipedia Asexternal Corpus", *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, ACM, New York, NY, USA, p. 797-798, 2007.

Macdonald C., Ounis I., "Expertise Drift and Query Expansion in Expert Search", *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, ACM, New York, NY, USA, p. 341-350, 2007.

Sahami M., Heilman T. D., "A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets", *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, ACM, New York, NY, USA, p. 377-386, 2006.

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

Voorhees E. M., "The TREC robust retrieval track", *SIGIR Forum*, 2005.

Wikipedia, "Wikipedia:Article titles — Wikipedia, The Free Encyclopedia", 2013.

Xu Y., Jones G. J., Wang B., "Query dependent pseudo-relevance feedback based on wikipedia", SIGIR '09, Boston, MA, USA, p. 59-66, 2009.

Xu Z., Akella R., "A Bayesian Logistic Regression Model for Active Relevance Feedback", *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, ACM, New York, NY, USA, p. 227-234, 2008.

Zhai C., Lafferty J., "A Study of Smoothing Methods for Language Models Applied to Information Retrieval", *ACM Trans. Inf. Syst.*, vol. 22, n° 2, p. 179-214, April, 2004.