International Journal of Database Theory and Application
Vol.8, No.3 (2015), pp.27-40
http://dx.doi.org/10.14257/ijdta.2015.8.3.03

Improving Query Expansion for Information Retrieval Using Wikipedia

Lixin Gan[1]and Huan Hong[2]

[1]*School of Math and Computer Science,JiangxiScience &Technology Normal University,Nanchang, China*
[2]*School of Computer Information Engineering,Jiangxi Normal University,Nanchang, China*
spiderganxin@163.com, honghuan252008@126.com

Abstract

Query expansion (QE) is one of the key technologies to improve retrieval efficiency. Many studies on query expansion with relationships from single local corpus suffer from two problems resulting in low retrieval performance: term relationships are limited and unlisted query terms have no expansion terms. To address these problems, relationships between terms captured from Wikipedia are superimposed to the basic Markov network that pre-built using single local corpus. A new larger Markov network is formed with more and richer relationship for each term. Evaluation is performed on three standard information retrieval corpuses including ADI, CISI and CACM.Experimental results show that the proposed technique of superimposed Markov network is effective to select more and confident candidatesfor query expansion and it outperforms other state-of-the-art QE methods.

Keywords: *query expansion; information retrieval; Wikipedia; Markov network*

1. Introduction

With the raid development of computer science and Internet, big data has brought tremendous challenges for information processing. It makes people difficult to search information they really need. Term mismatch is one of the fundamental challenges in web search, where a query and its relevant documents are often composed using different vocabularies and language styles. Therefore, query expansion (QE) is an effective strategy to address the challenge. Query expansion is the process of reformulating a seed query to improve retrieval performance in information retrieval operations by adding expansion terms related to original query terms, so that more relevant documents can be retrieved. QE is a long-standing research topic in information retrieval (IR) and more and more studies have been focus on it [1-10].Typical approaches of query expansion can be roughly divided into two categories: one is based on automatic relevance feedback (such as explicit feedback and pseudo relevance feedback (PRF)) [3-6] and the other is log-based QE which capture the correlation between query terms and document term from click data [7-10].

Query expansion methods above have been proved to be useful for improving the performance of IR. However, these methods only consider relationships between terms in a single local corpus in the process of query expansion. In fact, because the size of a single local corpus is relatively small, many information retrieval models suffer much from two problems in query expansion resulting in low retrieval performance: (1) Term relationships are limited and some terms may be false expansion terms and result in topic drift, although they have high relationships with original query terms solely captured from a single local corpus;(2) In particular, there are many terms onlyexisting in the query

ISSN: 2005-4270 IJDTA

setbut not inthedocument set for a given corpus. Such terms are named as unlisted query terms in this paper.

Table 1. Statistics: Unlisted Query Terms Information in Three Local Corpuses

Corpus	ADI	CISI	CACM
domain	information science	library science	computer science
#unlisted query terms	32	107	42
#effectiveunlisted query terms	32	38	42
# effective queries containing unlisted query term	17	21	25
# queries in corpus	35	112	64
#effective queries	35	76	64
%effective queries containing unlisted query terms to effective queries	48.57%	27.63%	39.06%

Table 1 above gives the statistics of unlistedquery terms in three standard information retrieval corpuses including ADI, CISI and CACM called local corpuses. In the local corpus, if a query in query set has relevance documents in the document set for retrieval, we regard it as an effective query. It shows that CISI corpus has 76 effective queries and 36 ineffective queries, while other corpuses such as ADI and CACM just contain effective queries. Because ineffective queries with no relevance documents in corpus don't take effect for retrieval, we only take effective queries into considered in this paper. Therefore, unlisted query terms contain two types: (1)effective unlisted query terms which appear in effective queries, (2) ineffective unlisted query terms appear in ineffective queries. In this paper, we focus on effective unlisted query terms and the ineffective unlisted query terms are removed in our work. From Table 1, it also shows that in a given local corpus, there are many effective unlisted query terms appearing in effective queries. Effective queries containing unlisted query terms account for 48% of total effective queries in ADI. At present, the unlisted words remains affect the efficiency of information retrieval.

Therefore, to solve the problem above, Wikipedia isutilized to help capturing more relationships between terms in our work. As a Web 2.0 knowledge system with the characteristics of open and user collaborative editing, Wikipedia has the following significant features: wide knowledge coverage, rich semantic knowledge, highly structured, rapidly speed of information update. Therefore, it is an ideal data resource for information retrieval [11-13]. Elsas applied the link structure of Wikipedia to query expansionin the context of the TRECBlog track, which can enhance blog feedback search task [11]. However, query dependent knowledge is not taken into consideration by the thesaurus[11].Xu applied Wikipedia resources in relevance feedback to prove the ways which get pseudo-related feedback from Wikipedia entity page and carry out extended terms selection superior to the basic model [12]. However, we are interested in selecting those Wikipedia articleswhich are related to query domain and use those to extract more term relationships. Y.Li made use of the categories in Wikipedia papers to carry out query expansion using assign [13]. The method shows improvement over PRF in measures favoring weak queries.

Different from above studies that utilize Wikipedia information for pseudo-related feedback, our work mainly focuses on the text content of Wikipedia and combine it with the local domain corpus. Term relationships extracted from Wikipedia are superimposed to the basic Markov network that pre-built using a single local domain corpus. Therefore, it makes Markov network with more and richer semantic information. Unlisted query terms will get candidates and are helpful for query expansion. In additional, it also updates the weight of relationships for listed terms and helps them have more confident

related terms. The proposed technique of superimposed Markov network is benefit to select candidates for unlisted query terms as well as listed terms in query expansion and it outperforms other state-of-the-art QE methods.

2. Related Work

2.1 Query Expansion

The purpose of information retrieval is to search relevance documents to query in the document set. Given a query Q and a document D, the basic idea behind information retrieval model is to compute the conditional probability $P(D|Q)$.The documents are ranked in descending order of this probability.Assuming that terms in the query are independent, we have a general model formulated as follows:

$$p(D|Q) \propto \sum_{q_i \in Q} p(q_i|Q)\, p(q_i|D) \tag{1}$$

We can observe that the formula above still requires query terms to appear in a document for retrieval. However, in reality, there often exits a problem that is term mismatch between queries and documents. The problem of term mismatch occurs because people often use different terms to describe concepts in their queries from those to describe the same concepts in their documents. Query expansion has been suggested as a technique for dealing with this problem [14].

With respect to formula (1), query expansion consists of finding abetter way of estimating$P$ ($q_i|D$), so that not only the termsexpressed in the query will have a non-zero probability, but also have other related terms. Therefore, the basic query expansion model is shown as follows:

$$P(q_i \mid D) = (1-\lambda)P(q_i \mid D) + \lambda \sum_{t_j \in V} W(t_j, q_i) P(t_j \mid D) \tag{2}$$

Where$V$ is the vocabulary of the corpus and λ is a smooth parameter.

Putting it into formula (1), we obtainthe following query expansion formula:

$$P(D \mid Q) = \sum_{q_i \in Q} ((1-\lambda)P(q_i \mid D) + \lambda \sum_{t_j \in V} W(t_j, q_i) P(t_j \mid D)) \times P(q_i \mid Q) \tag{3}$$

Where $t_k$is a term related with q_i.

Noted that formulas (2) and (3) still require query terms has related termsin the vocabulary of the corpus. If a query term $q_i$is an unlisted term in corpus, it means that q_i solely appears in query set but not in document set and $q_i$has no related terms in corpus. Then both P ($q_i|D$) and W (t_k,q_i) are zero. Therefore, if a query contains unlisted query terms, query expansion doesn't work for such query terms effectively resulting in low retrieval performance. For example, given a query = {***apple***, ***software***, ***hardware***} which represents the query topic "***information about software and hardware of apple company***". If the query term "***apple***" is an unlisted term, there is no term relationship with it from corpus. Owing to the expansion of query term "***software***" and "***hardware***", search results maybe concentrate on information about software and hardware. Therefore, query expansion based on such method maybe leads to too much noisy and topic drift. In addition, as all we known, some terms such as "***Ipad***" and "***Iphone***" are related with "***apple***". Although some documents having such related terms are not necessarily retrieved becausethe unlisted term "***apple***" has no candidate in query expansion.Therefore, in this paper, we focus on the query expansion problem where the

goal is to extract related terms for unlisted query terms as well as listed query terms using Wikipedia.

2.2 Markov Network Model

With respect to formulas (3), it is important to determine$W$ (t_k,q_i). In this paper, we develop information retrieval model in a unified framework-Markov network which canmodel term relationships and information retrieval model to explore the impact ofrelevance information to retrieval performance. The Markov network representationmodel can model arbitrary features including term relationship and all kinds of termfeatures [15].This work extends previous work [16]by adding term relationship from Wikipedia to constructing a new Markov semantic network. We first describe the Markov network information model in more detail and then present ourextensions and modifications.

The Markov network iscapable of efficiently representing relevance in knowledge and is easily gotten from training data with strong learning and inferring capability[16]. It can be used to represent any of the classic models in IR. A Markov network is an undirected graph G and is expressed by G (V, E,W). Let V be the set of term nodes and E be the set of undirected edgesin the graphrespectively, and W be the set of weight value between terms.In particular, a term in the graph is independent of its non-neighbors given observed values for its neighbors. The Markov network is shown as Figure 1.

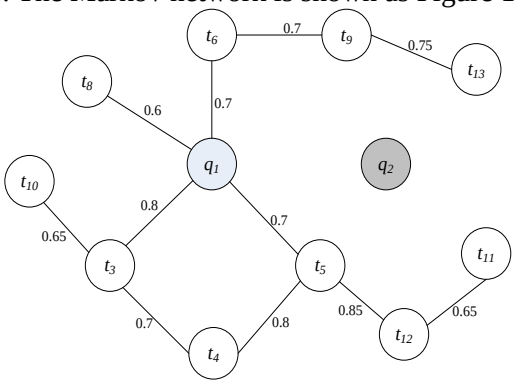


Figure1.B-MarkovNetwork from Local Domain Corpus

3. TheThree-stepConstructionof MarkovNetwork

The Markov network is built according to term relationships. In our work, we combine two types of term relationships from single local domain corpus and Wikipedia corpus, respectively. The construction of Markov network in our work is an extension of the work [16]. The construction of Markov network takes three steps as follows: 1) build Markov network using the single local corpus asa basic Markov network called B-Markov network; 2) build Markov network using Wikipedia corpus called W-Markov network; 3) W-network is superimposed to B-Markov network so as to form a larger Markov network called C-Markov network with more and richer relationship for each term.

3.1 Construction of B-Markov Network

The construction of B-Markov network from the single local corpus is similar to previous work [16]. First of all, term correlativity can be measured by mutual information (MI), latent semantic index or term co-occurrences. Considered undirected characteristics of Markov network, our work simply adopt co-occurrences between terms to measure term relationship as follows:

$$W_B(t_i,t_j) = \frac{N(t_i,t_j)}{N(t_i)+N(t_j)-N(t_i,t_j)} \quad (4)$$

Where $W_B(t_i,t_j)$ measures the relationship between $t_i$and$t_j$in the B-Markov network.$N(t_i,t_j)$is the frequency of co-occurrences of $t_i$and t_j in the document set for a local corpus. $N(t_i)$and $N(t_j)$ have the definition similar to$N(t_i,t_j)$.

In this paper, our work is based on the hypothesis that "if there is a relationship between terms, there exists an edge between them". For example, given a query Q ={q_1,q_2}, q_1 is a term existing in the local corpus with a related term set $S_B(q_1)$ ={t_3, t_5,t_6, t_8}, while q_2 is an unlisted term not in the document set of the local corpus and it has an empty related term set $S_B(q_2)$={}. Therefore, the B-Markov network of Q is shown as Figure 1 above. The edge represents that two terms are related and the number on the edge is the weight value of their relationship. From Figure 1 above, we know that if a local corpus contains unlisted query terms, the B-Markov network consists of two types of nodes: (1) terms appear in documents set and have edges with its related terms like q_1; (2) unlisted terms solely appear in query set but not in any document as single nodes in B-Markov network like q_2.

3.2 Constructionof W-Markov Network

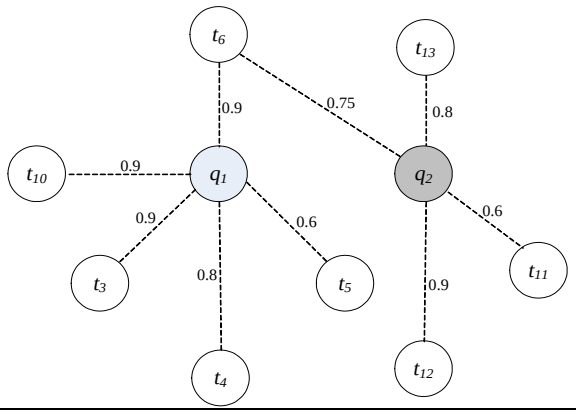
From Figure 1, we know that q_2 has no related term. When expanding such unlisted query term like q_2, its candidate term set is empty and query expansion can't take effect for it. Therefore, to solve the problem above, we consider Wikipedia as an additional corpus. In order to capture high quality of related terms and to avoid too much noisy, we download the text content of entity pages from Wikipedia English site according to the categories of pages which are related to the domain of the local corpus. Therefore, Wikipedia corpus is built consistent with the domain characteristics of the local corpus as an additional corpus. In order to capture more high ranking expansion terms and to reduce computing cost, we only retain terms in Wikipedia corpus which also appear in the local corpus. It means that each term nodet_i in W-Mark satisfies the condition $t_i \in (V_W \cap V_B)$),where V_W and V_B are the node sets in W-Mark and B-Mark respectively.

We also adopt the same approach above to compute relationships between terms $t_i$and t_j in Wikipedia corpus, called $W_W(t_i,t_j)$ .The work for constructing W-Markov network using Wikipedia is similar to B-Markov network. Therefore, we can get W-Markov network from Wikipedia corpus. For the example Q = {q_1, q_2}above, W-Markov network of the query Q is shown as Figure 2. In Wikipedia corpus, query term q_1 and q_2 are both appear and have their related terms respectively. The related term set are $S_W(q_1)$ = {t_3, t_4, t_5, t_6,t_{10}} and $S_W(q_2)$ = {t_6,t_{11},t_{12},t_{13}} respectively. Therefore, compared to B-Markov network, q_2 is not an unlisted query term and can obtain some related terms from W-Markov network. In addition, the query term q_1 also get more additional related terms from W-Markov network such as $t_3$and $t_4$which have no relationship with $q_1$in B-Markov.

3.3 Superimposition to C-Markov Network

In this paper, we combine W-Markov networkwith B-Markov networkto form a new larger Markov network-C-Markov network with more and richer relationships between terms. Our approach is also related to the work [17]which uses the technique of superimposition tomine hidden relationships into graphs. And our construction of C-Markov network is inspired by this work. Once the W-Markov network is created, we cansuperimpose it against the B-Markov network. The superimposition processing contains two components: edge superimpositionand weight superimposition. The construction for C-Markov network is shown in our superimpositionalgorithm as Algorithm 1.

Figure 2. W-Markov Network from Wikipedia Corpus



Algorithm 1:superimposition for C-Markov network

Input: *G (V_B, E_B,W_B) in B-Markov network and G (V_w, E_w, W_w) in W-Markov.*
Output: *G (V_c, E_c,W_c)in C-Markov network.*
Initialization: *V_c= V_B,E_C=E_B, W_C=W_B.*
forall *term t_i($\in V_B$) do*
  if $E(t_i, t_j) \in E_B \wedge E(t_i, t_j) \in E_w$, then $W_C(t_i, t_j)=max(W_B(t_i, t_j), W_w(t_i, t_j))$;
~~else~~if $E(t_i, t_j) \notin E_B \wedge E(t_i, t_j) \in E_w$, then add an edge $E(t_i, t_j)$ into E_c;$W_C(t_i, t_j)=W_w(t_i, t_j)$;
Return *G (V_c, E_c, W_c).*

According to the superimposition algorithm above, the C-Markov network is constructed easily. For instance, the W-Markov network for a given query $Q = \{q_1, q_2\}$ shown in Figure 2 is superimposed to its B-Markov network in Figure 1 forming a new C-Markov network in Figure 3.

As shown in Figure 3, since the motivation of this work is to extract more and richer related terms for query expansion, the combined term set is built on the vocabulary of the local corpus.After node superimposition, terms of $V_C$in C-Markov networkare as same as that of $V_B$inB-Markov network.Due to edge superimposition, there are three types of edges in Figure 3represented by different line shapes.Thefinesolid edge represents term relationshipssolely extracted from the single localcorpus such as these edges{$E(q_1, t_8)$, $E(t_6, t_9)$, $E(t_9, t_{13})$, $E(t_3, t_{10})$, $E(t_3, t_4)$, $E(t_5, t_{12})$,$E(t_{11},t_{12})$}.While the dotted edges represent term relationshipssolely from Wikipedia corpus suchas {$E(q_1, t_4)$,$E(q_1,t_{10})$, $E(q_2, t_6)$,$E(q_2, t_{11})$, $E(q_2, t_{12})$,$E(q_2, t_{13})$}.

We notice that there are bold solid edges due to the superimposition of term relationships in both corpuses such as {$E(q_1, t_3)$, $E(q_1, t_5)$, $E(q_1, t_6)$}.

In the process of weight superimposition, we adopt the maximumweight of term correlativitybetween two corpuses as their final term correlativity as$W_C(t_i,t_j)$expressed as:

$$W_C(t_i,t_j) = MAX\{W_B(t_i,t_j), W_W(t_i,t_j)\} \quad (2)$$

That is, the term weight$W_C(t_i, t_j)$in C-Markov network is shown as:

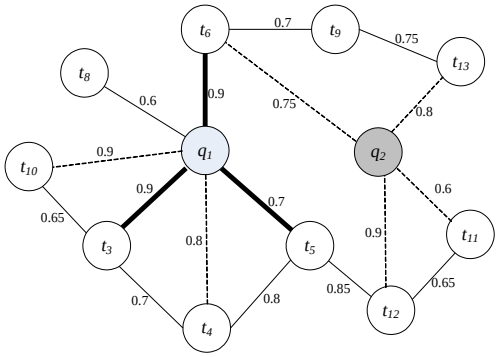


Figure 3. C-MarkovNetworkwhere W-Markov network is Superimpose to B-MarkovNetwork

$$W_C(t_i,t_j)=\begin{cases} W_B(t_i,t_j) & \text{if term relationship only from local coprus} \\ W_W(t_i,t_j) & \text{if term relationship only from Wikipedia coprus} \\ MAX\{W_B(t_i,t_j),W_W(t_i,t_j)\} & \text{if term relationship from both copruses} \end{cases} \quad (3)$$

Therefore, in C-Markov network shown as Figure 3, the weights on the fine solid edges all come from $W_B(t_i,t_j)$ in the B-Markov network. The weights on the dotted edges are only extracted from C-Markov network. While the weights on the bold solid edges lies on the maximum value between them in both corpuses. Therefore, weight values of $W_C(t_i,t_j)$ on the bold solid edges in the Figure 3 are shown in Table 2.

Table 2.Weight Value on the Dotted Edges in Figure3

Related term pair	W_C	W_B	W_W
(q_1,t_3)	0.9	0.8	0.9
(q_1,t_5)	0.7	0.7	0.6
(q_1,t_6)	0.9	0.9	0.7

After theC-Markov network of term relationship is built, the next step is query expansion. Query expansion aims to generate additional expansion terms that are statistically related to original query terms. Therefore, the core issue of query expansion technique is how to design and utilizethe sources of related terms. For each term t_i, its term relationship with other terms can be easily extracted from C-Markov network. Since we try to improve query expansion by getting more candidates for originalquery terms and reducing noisy, the possible expansion terms should be quantified. In this paper, we select no more than 50 related terms for each query term as candidates in the process of query expansion. We rank related terms for a given term based on two hypothesis that: (1) the higherweight with an original query term, the more important the term is, and it has more chance to be an candidate for query expansion; (2) For a given term, if its term relationships from the local corpus and the Wikipedia corpus have the same weight to it , we think term relationships from the local corpus are more important than those from the Wikipedia corpus and rank term relationshipfrom the local corpus prior to that from Wikipedia corpus. Therefore, for each term t_i, the list of related terms for query expansionas $L(t_i)$ is sorted in descending orderbased on the hypothesis above. From the C-Markov network in Figure3, we note that the network is larger and contains more rich semantic information than any single network in Figure 1 and Figure 2.For instant, in Figure 3, the list of related terms for each query term is:$L(q_1)$={t_3,t_6,t_{10},t_4,t_5, t_8}and $L(q_2)$ ={ t_{12}, t_{13}, t_6, t_{11}}. While the B-Markov network from a single local corpus in Figure 1, the

list of candidates for each query term is: $L(q_1) = \{t_3, t_6, t_5, t_8\}andL(q_2)=\{\}$.We notice that since W-Markov network is superimposed to B-Markov network, C-Markov network contains more term relationships not only for unlisted query terms but also for general listed terms and is benefit to improve the efficiency of query expansion.

4. Experiment Validation

4.1 Experimental Setup

For validating our proposal, we apply it to threestandard corpuses in information retrieval combined with Wikipedia corpuses. There are ADI, CISIand CACM as local corpuses. We gathered web information from Wikipedia as Wikipedia corpuses according to the domains of local corpuses.

Local corpuses: The local corpuses-ADI, CISI and CACM – are described in the Table 1 above and Table 3.

Table 3. Summary of Local Corpuses

Local corpus	ADI	CISI	CACM
Domain	Information Science	Library Science	Computer Science
#terms in corpus	925	5601	5083
#documents	82	1460	3024
#effective queries	35	76	64
#final effective unlisted query terms	32	35	30

From CISI local corpus in Table 1 above and Table 3, there are 36 queries with no relevance documents in corpus. In retrieval processing, they bring too much noisy results leading to low precision. Therefore, we remove them and only retain effective queries for retrieval in our experiments.Since the motivation of this work is to improve query expansion, we only focus on effectiveunlisted query termsand extract more relationship for them from Wikipedia corpus.

In local corpuses, there are some spelling mistakes. In CISI local corpus, there are 5 terms with spelling mistake such as "***prospct***", "***abstreact***", "***analysins***", "***compuyter***" and "***suybsystem***". And there is also 1 spelling mistake term in CACM. We have manually corrected them. From CACM local corpus, we manually remove 12 terms from the set of effective unlisted query terms because they are people' names such as "***GerardSalton***". But in the processing of stemmer, they are stemmed as independent terms such as "***Gerard***" and "***Salton***". If expanding such independent terms of people' names, they will lead to too much noisy. One of the methods used to tackle this problem is to identify people' name as a term. But it refers to other research area on named entity recognition and resolution. In our future work, we will pay attention to it. Therefore, we get the set of final effective unlisted query terms in our experiments.

Wikipedia corpuses: In order to capture high quality of related terms and to avoid too much noisy, we download the text content of entity pages from Wikipedia English site according to the categories of pages which are related to the domain of the local corpus. For example, in order to setup an additional corpus from Wikipedia for CISI, we download one of entity pages such as "***Library and information science***" form Wikipedia.[1] Its categories contain "***informationscience***" "***librarians***" and "***libraryscience***" which are related to the domain "***LibraryScience***" of CISI corpus. We remove all web labels in entity pages and save each page as a document. We setup three additional corpuses from

[1]http://en.wikipedia.org/wiki/Library_and_information_science.

Wikipedia for local corpuses named as Wikipedia corpuses. The information of three Wikipedia corpuses is shown in Table 4.

Table 4.Summary of Wikipedia Corpuses

Wikipedia corpus	WIKI_ADI	WIKI_CISI	WIKI_CACM
#documents	11	29	22
#terms in corpus	3816	5511	3172
#terms in local corpus	699	2768	1762
#listed terms in local corpus	675	2733	1738
#unlisted terms in local corpus	24	35	24

In Table 4 above, WIKI_ADI, WIKI_CISI and WIKI_CACM are additional corpuses for local corpuses ADI, CISIand CACM respectively. Although the size of each Wikipedia corpus is relatively small, the total number of terms in each Wikipedia corpus is larger relative to the corresponding local corpus. That is because each document in Wikipedia corpus is longer and contains more and richer information. Our work focuses on extracting more and richer relationship between terms which appear in local corpus for query expansion. Therefore, for Wikipedia corpus, we retain these terms which also appear in local corpus as "*terms in local corpus*". These terms consist of the node set $V_W$in W-Mark.We name terms in both Wikipedia corpus and local corpus as "*listed terms in local corpus*". Common terms in Wikipedia corpus reach an half of terms in local corpus on average. The relationships for such terms are expressed as bold solid edges in C_Markov. Terms in Wikipedia corpuscovers unlisted termsof local corpus more than 75%, especially up to 100% in CISI. This type of term relationship is shown as the dotted edges in C_Markov.

Text Preprocess: All documents in both local corpusesand Wikipedia corpuses have been processed in a standard manner: only titles and bodies in document are used, terms are stemmed using the Porter Stemmer, stop words are removed and words are converted into lowercase.

Evaluation Metrics: The experimental results are measured using 11-Avg (This precision versus recall curve is based on 11 standard recall level which are 0%, 10%, 20%, … , 100%.) and 3-Avg (This precision versus recall curve is based on 11 standard recall level which are 20%, 50% and 80%.).

4.2 Experimental Results

In order to evaluate the retrieval performance of our proposal, we compare our query expansion models to the baseline model. Our motivation of this work is to evaluate the retrieval performance according to query expansion using term relationship from Wikipedia. Therefore, we simply adopt the traditional query expansion technology in our query expansion models based on an independent assumption of query terms. Expansion terms are selected as candidates for query expansion according to their high ranking relationships with query terms. The key difference of our two query expansion model is that relationships of expansion terms come from different date sources.

Baseline Model:We use the classical unigram model without any expansion as our baseline model.

Query expansion + local corpus:It is a query expansion model that its term relationships are solely captured from the single local corpus in [16].

Query expansion + local corpus and Wikipedia corpus: The query expansion model consider the combined term relationship both from the local corpus and Wikipedia corpus.

Table 5 and Table 6show the retrieval performance on 11-Avg and 3-Avg respectively. The percentages in the table are relative changes in respect to the baseline model.

Table 5.11_AVG Results on Three Corpuses

Model \ Corpus		ADI	CISI	CACM
Baseline model		42.0%	14.2%	23.0%
Query expansion	+local corpus	45.6%	21.9%	31.4%
	+local corpus and Wikipediacorpus	**49.3%**	**24.9%**	**35.4%**

Table 6.3_AVG Results on Three Corpuses

Corpus \ Model		ADI	CISI	CACM
Baseline model		42.0%	17.9%	23.0%
Query expansion	+local corpus	45.4%	20.5%	30.1%
	+local corpus and Wikipedia corpus	**51.1%**	**22.2%**	**33.6%**

Table 5 and Table 6show the contribution of employing combined term relationships from both local corpus and Wikipedia corpus on three standard datasets. We can find that both query expansion models outperform the baseline model in all corpuses. They show that query expansion is benefit to improve retrieval effect.

As we can see from Table 5and Table 6, Our query expansion model with superimposed term relationships both from the local corpus and Wikipedia corpus performances the best both on 11-Avg and 3-Avg in all test corpuses. Our approach dramatically enhances 11_avg measure and 3_avg measure by 11.6% and 12.4% relative to the baseline model in CACM corpus. It proves that our proposal is helpful for improving query expansion due to more and richer term relationship extracted from C_Network.Thetechnique of superimposed Markov network not only helps listed query terms get more confident candidates, but also helps unlisted query terms participate in query expansion benefitting from its term relationships captured in W_Network.

As a complementing technique, comparing to the query expansion model with term relationships solely from the local corpus, our approach obtains the best performance in 11_AVG for ADI, increases of 4%. The reason is that it is benefit from many common terms both in the combined network strengtheningtheir term relationship.WIKI_ADI corpus has 699 common terms appearing in local ADI corpus,occupying up to 90% in ADI. Although WIKI_CACM corpus contains only 30% common terms in localCACM corpus, our proposal performances the best in 3_AVG for CACM, with increasing of 5.7% benefitting from adding related candidates into query expansion for unlisted terms. From Table 4 above,WIKI_CISI corpus extracts term relationships for all unlisted terms in local CISI corpus. However, our proposal has the least improvement 3% and 1.7% on 11_AVG and 3_AVG respectively for CISI. We find that there are a few candidates added in query expansion relative to other two corpuses because it contains many too long queries. Strengthen Relationships for listed terms from Wikipedia corpus have little chance to work for improving performancein query expansion. For example, the NO.39 of queries in CISI is: "*The progress of information retrieval presents problems of maladjustment and dislocation of personnel. Training and retraining of people to use the new equipment is important at all levels. Librarians, assistants, technicians, students, researchers, and even executives will need education to learn the purpose, values, and uses of information systems and hardware. What programs have been developed to change the attitudes and skills of traditional workers and help them to learn the newer techniques?*".Therefore, the

improvements for CISI mainly profit from adding candidates of unlisted query terms using Wikipedia corpus.

With respect to smoothing parameter λ involved in query expansion, we take empirical measures. For the variant, we set the parameter to vary from 0 to 1. Figure 4-6 show the effect on 11_AVG and 3_AVG of varying thesmoothing parameter λ on in these three corpuses. From experiments, we can find that the whole tendency of smoothing parameter λ in all corpuses is same and its value is very small, between (0.1-0.2).It shows that query expansion is a supplement technology for improving retrievalperformance.

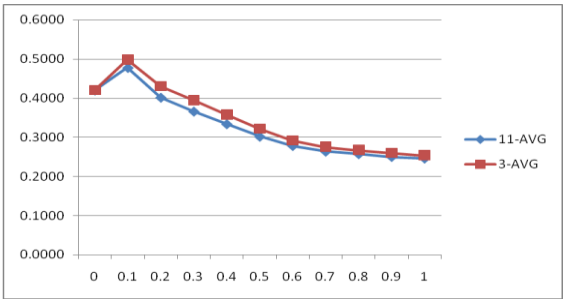


Figure 4. Effect of Smoothing Parameter λ on ADI Corpus

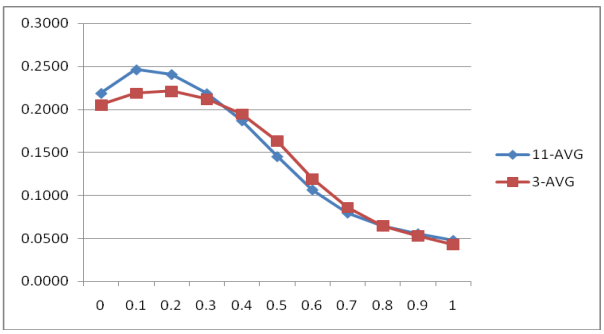


Figure 5. Effect of Smoothing Parameter λ on CISI Corpus

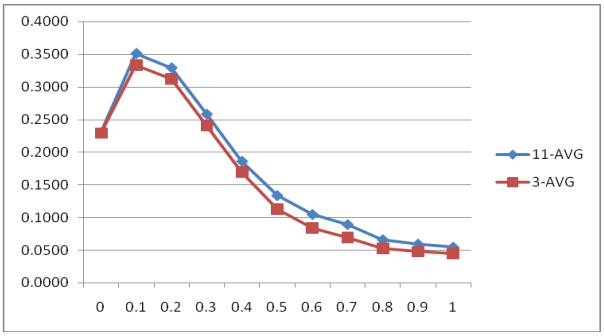


Figure 6. Effect of Smoothing Parameter λ on CACM Corpus

5. Conclusion

Due to the small size of the single local corpus, many information retrieval models suffer from two problems in query expansion resulting in low retrieval performance: term relationships of listed terms are limited and unlisted terms have no expansion terms. To solve the problems above, we present a new approach to extract more term relationship from Markov network for query expansion. In this

paper, term relationship extracted from Wikipedia corpus is superimposed to the basic Markov network that pre-built using the single local corpus. Therefore, a new larger Markov network is built with more and richer term relationship for unlisted terms as well as listed terms. Evaluation is performed on three standard information retrieval corpuses including ADI, CISI and CACM. Experimental results show that the proposed technique of superimposed Markov network is effective to select more and confident candidates for query expansion and it outperforms other state-of-the-art QE methods.

In future work, we will focus on getting more high ranked candidates for query expansion by relieving the independence between query terms. In order to avoid the problem that people' name is stemmed as independent terms, we will apply technologies of named entity recognition and resolution into text preprocess. An appealing direction would be to integrate the structured information from Wikipedia to improve query expansion.

Acknowledgements

This work was supported in part by Natural Science Foundation of Jiangxi Province under grant to NO.20122BAB21103 and NO.00029511101228076, Humanities and Social Sciences Foundation of Jiangxi provincial universities grant to NO.JC1312 and JD1164, Education Reform Project of Jiangxi universities grant to NO.JXJG-13-10-13.

References

[1] K. Tamsin Maxwell and W. B. Croft.Compact Query Term Selection Using Topically Related Text. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval,**(2013)**July28-August01;Dublin, Ireland.
[2] Y. Lin, H. Lin and S.Jin, "Social Annotation in Query Expansion: a Machine Learning Approach",Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval,**(2011)**July24-28; Beijing, China.
[3] D. Metzler and W. B. Croft, "Latent concept expansion usingMarkov random fields",Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, **(2007)**July 23-27; Amsterdam, Holland.
[4] G.Cao, J-Y.Nie, J.Gao, and Robertson, "S. Selecting good expansion terms for pseudo-relevance feedback",Proceedings of the 31st annual international ACM SIGIR conference onResearch and development in information retrieval, **(2008)**Jul 20-24; Singapore.
[5] V.Lavrenkoand B.Croft,"Relevance-based language models",Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, **(2001)**September9-12; New Orleans, LA.
[6] C.Zhaiand J.Lafferty,"Model-based feedback in the kl-divergence retrieval model",Proceedings of the 10th International Conference on Information and Knowledge Management,**(2001)** November 05-10; Atlanta, Georgia, USA.
[7] J. Gao, Gu Xu and Jinxi Xu, "Using Query Expansion Using Path-Constrained Random Walks",Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, **(2013)** July 28-August 01; Dublin, Ireland.
[8] J. Gaoand J-Y.Nie, "Towards concept-based translation models using search logs for query expansion", Proceedings of the 21st ACM international conference on Information and knowledge management, **(2012)** October 29 - November 02; Maui, HI, USA.
[9] J. Gao, S.Xieand X.He,"A. Learning lexicon models from search logs for query expansion", Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, **(2012)** July 12-14, Jeju Island, Korea.
[10] S.Riezlerand Y. J.Liu, "Query rewriting using monolingual statistical machine translation",Computational Linguistics, vol. 3, no. 36,**(2010).**
[11] J. L. Elsas, J. Arguello, J. Callan and J. G.Carbonell, "Retrieval and feedback models for blog feed search", Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, **(2008)**July 20-24; Singapore.

[12] Y. Xu, G. J. F. Jones and B. Wang, "Query dependent pseudo-relevance feedback based on Wikipedia",Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, **(2009)**July 19-23;Boston, Massachusetts.
[13] Y. Li, W. P. R. Luk, K. S. E. Ho and F. L. K. Chung, "Improving weak ad-hoc queries using Wikipedia as external corpus", Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,**(2007)**July23-27;Amsterdam, Holland.
[14] B. Croft and J. Lafferty, "Language Models for Information Retrieval", Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, **(2005)** August 15-19; Salvador, Brazil.
[15] J.Zuo, "Research on Markov Graph Model in Information Retrieval", Doctoral Dissertation, Jiangxi University of Finance and Economics, Nanchang, **(2011)**.
[16] L.Gan,"The Information Retrieval Model based on Markov concept", Jiangxi normal University,Nanchang,**(2007)**.
[17] B. W. On, E. Elmacioglu, D. Lee, J. Kang and J.Pei,"Improving grouped-entity resolution using quasi-cliques", Proceedings of the Sixth International Conference on Data Mining (ICDM),**(2006)**December 18-22;Hong Kong, China.

Authors

LixinGan, she was born in 1982. She is a Ph.D. candidate and a lecturer at Jiangxi Science & Technology Normal University. Her research interests include information retrieval, information extraction and data mining.

Huan Hong, he was born in 1991. He is a Postgraduate at Jiangxi Normal University. His research interests include information retrieval,NLP and data mining.

