The Electronic Library

Inverse local context analysis: A method for exhaustively gathering documents from limited accessible data sources

Wei Lu, Xinghu Yue, Qikai Cheng, Rui Meng,

Inverse local context analysis

A method for exhaustively gathering documents from limited accessible data sources

Wei Lu, Xinghu Yue, Qikai Cheng and Rui Meng
School of Information Management, Wuhan University, Wuhan, China

Received 10 December 2014
Revised 5 April 2015
9 April 2015
Accepted 25 May 2015

Abstract

Purpose – The purpose of this paper is to explore the use of inverse local context analysis (ILCA) to obtain data from limited accessible data sources.

Design/methodology/approach – The experimental results show that the method the authors proposed can obtain all retrieved documents from the limited accessible data source using the least number of queries.

Findings – The experimental results show that the method we proposed can obtain all retrieved documents from the limited accessible data source using the least number of queries.

Originality/value – To the best of the authors' knowledge, this paper provides the first attempt to gather all the retrieved documents from limited accessible data source, and the efficiency and ease of implementation of the proposed solution make it feasible for practical applications. The method the authors proposed can also benefit the construction of web corpus.

Keywords Query expansion, Limited accessible data source, Local context analysis, Total recall, Exhaustive search

Paper type Research paper

Introduction

The World Wide Web provides a large and heterogeneous data corpus for scholars from different domain areas to investigate diverse research questions (Robert, 2009). It has been utilized as a language corpus for linguists (Baroni and Bernardini, 2004; Kilgarriff and Grefenstette, 2003) and successfully applied in many natural language processing (NLP) applications, such as machine translation (Grefenstette, 1999), term extraction and grammar checking (Liu and Curran, 2006). In information retrieval, web data have been successfully applied in web page classification (Qi and Davison, 2009), image retrieval and annotation (Wang *et al.*, 2008) and user query classification (Hu *et al.*, 2009).

Search engines and online databases provide a convenient way of accessing these resources (Zhu and Xie, 2003). Many successful studies created corpora, partially or fully, by utilizing search engine results. Examples include public opinion monitoring and trend analysis (Sharoff, 2006; Wang, 2011; Zhu, 2012). When creating corpora from search engines or online databases, a common method is to submit a query, manually or automatically, to the search system and then gather documents from the search results. However, to reduce system cost, modern commercial search engines, such as Google and Baidu (the largest Chinese search engine), usually do not provide all of the retrieved documents in their results. According to the statistics of the general search engines in 2014, for each query, Google returns less than 1,000 results and Baidu returns, at most,

The Electronic Library
Vol. 34 No. 3, 2016
pp. 405-418

0264-0473
DOI 10.1108/EL-12-2014-0211

760 results. This makes constructing a coverage-oriented data corpus problematic. The authors of this paper call the limited access to full search results the *access restriction problem*.

Constructing coverage-oriented data corpora is needed in many real scenarios. For example, researchers may need to find as many relevant publications as possible in their research domains for a better understanding of research trends (Datta *et al.*, 2008; Wilson, 2000) and news websites may need such information to create a collection of news about certain events (Dou *et al.*, 2012; Ku *et al.*, 2006).

Given a scenario where a librarian or system analyst wants to write an analysis report of a given technique area, he or she needs to find all of the relevant documents on the specific topic. In general, when an individual submits a query to Google Scholar and retrieves a limited number of retrieved documents, it is impossible for the user to gather all of the possible documents. A set of sub-queries sharing the same information need with the original query may be helpful in gathering more documents. However, without a good strategy for formulating sub-queries, the overlap among results makes it difficult to gather relevant documents efficiently from the limited accessible data sources.

Given that a limited size of results is retrieved for each query and similar queries may bring overlapping results, the authors propose a novel technique to handle this problem. The assumption is that a coverage-oriented data corpus can be divided into nearly mutually exclusive sub-corpora, where each corpus can be retrieved by a single *optimal* query. However, such a query may not exist in reality. With this new technique, the problem can be converted into finding a set of optimal queries that are close enough to answer the same information need expressed by the original query. In the meantime, these queries can also separate the data corpus into nearly mutually exclusive sub-corpora.

The most closely related work from previous studies is query expansion (Chum *et al.*, 2007; Qiu and Frei, 1993). However, the research goal is significantly different from query expansion problems in traditional information retrieval, as query expansion is usually adopted to find the most relevant documents, whose purposes are mostly about the precision or the leverage between precision and recall, instead of solely on recall, or coverage as in this research.

In this paper, the authors propose a coverage-oriented query expansion method based on local context analysis (LCA). LCA has been used in information retrieval for a long time. Traditionally, LCA involves expanding a query using terms co-occurring with more query terms over other terms. However, for the purpose of obtaining all relevant documents from a limited accessible data source, the authors use the idea of LCA and select terms co-occurring with fewer query terms. In the experiments, the performance of the proposed method is compared with other expansion methods in exhausting all the relevant documents from a limited accessible data source. Experimental results show that the method proposed in this paper can exhaust most of the relevant documents from the limited accessible data source with the least number of queries.

Literature review

After a literature search, the authors conclude that there is no related research focusing on how to gather all retrieved documents from a limited accessible data source. However, there is a high correlation between this current work and the query expansion

method in information retrieval. *Query expansion* is the process, techniques, algorithms or methodologies that reformulate the original query to improve retrieval performance in information retrieval. In the context of information retrieval, query expansion involves evaluating the original query and reformulating it to match additional documents which are relevant to the information need.

The query expansion method was first introduced in text retrieval and became widely used in further research. Jones and Barber (1971) grouped words into clusters based on co-occurrences and then used the clusters for query expansion. However, a serious limitation that they noted was that the method was unable to handle ambiguous terms. If a query term has several meanings, then the term clustering-based method would reformulate the query by adding terms of various meanings and make the reformulated query even more ambiguous, thereby lowering retrieval performance. To address the word ambiguity problem, Jing and Croft (1994) proposed Phrasefinder. They exploited the mutual disambiguation of the query terms by selecting expansion terms based on their co-occurrence with all query terms. Terms co-occurring with more query terms are preferred over terms co-occurring with fewer query terms. Phrasefinder is one of the most successful query expansion methods, despite its problem in efficiency, as the creation of the pseudo-documents requires the co-occurrence data for all possible concept-term pairs. To improve the efficiency of query expansion, some researchers proposed using the top-ranked documents for query expansion (Cao *et al.*, 2008; Lee *et al.*, 2008; Parapar and Barreiro, 2011).

As LCA is simple and effective, it has become a widely used query expansion technique in information retrieval. Feng and Huang (2011) extracted terms relevant to the original query from n top-ranked retrieved local documents and then identified those featured terms as candidate expansion terms, according to the frequency of each featured term in the local documents and the correlation between each featured term and the original query terms. Their experiments showed that using the terms selected from top-ranked documents improved retrieval performance. Huang *et al.* (2006) used text classification and co-training techniques to identify relevant passages. The selected passages were used to expand query terms and re-formulate the probabilistic weighting function. With the utilization of machine learning methods, their results showed that the methods proposed were able to help with improving precision. One more conclusion of their results is that the performance of local analysis relies heavily on the performance of the original search results; if most of the top-ranked documents are not relevant, then the query expansion method's performance will be very erratic.

The query expansion method is also widely used in social media information retrieval. Social media, such as microblogs, mainly contain insufficient, temporal and non-standard text; thus, traditional retrieval methods based on keyword matching are not appropriate to be applied in the retrieval of social media information. By using the query expansion mechanism, however, the performance of social media information retrieval can be enhanced. Tang and Fang (2014) constructed a semantic retrieval model for microblog retrieval by combining semantic features, topics and text similarity features to select terms for query expansion. Their results showed that the query expansion method could improve retrieval performance of microblog retrieval. Biancalana and Micarelli (2009) utilized social tagging data, such as human-generated tags, annotations and user mark-up for performing personalized query expansion, and

concluded that their proposed approach significantly benefited personalized web search by leveraging the users' social media data.

In the field of multimedia information retrieval, multimedia objects, such as web images, videos and audio files, contain few text descriptions. Document-based query expansion techniques are thus difficult to apply. To solve this problem, Huang *et al.* (2001) presented a query expansion mechanism by expanding user queries through mining the associations among query terms in query session logs. In their experiment, the recall rate of web image retrieval was shown to be effectively increased. Crespo *et al.* (2012) utilized the hierarchical structure by which the MeSH descriptors are organized and selected the MeSH word ontology as the query expansion. They concluded that query expansion exploiting the hierarchical structure of the MeSH descriptors achieved a significant improvement in image retrieval systems.

In summary, research about query expansion has mainly focused on document-based retrieval, web multimedia retrieval and social media retrieval for improving the performance of information retrieval. The main idea of query expansion is to reformulate queries to retrieve additional results. The trade-off between precision and recall is a major problem with query expansion. While traditional query expansion application concentrated more about the precision of search results, in the current paper, the authors focus on utilizing the idea of query expansion to gather more results than can a single query. The objectives are different from previous research work: to obtain all the relevant documents from a limited accessible data source with the minimum number of queries and to improve the efficiency of data acquisition. The authors argue that query expansion can also be applied to obtain as many documents as possible from a limited accessible data source. As most search engines and online databases can only return a limited number of results for one query, using a single query is not capable of gathering all the results which are relevant to the input query. To gather all the results, the query should be expanded to generate a set of queries for retrieval.

Data acquisition method based on query expansion

Obtaining data from a limited accessible data source

Currently, limited accessible data sources, including most commercial search engines and scholarly databases, have become important sources for collecting data, which benefit corpus construction on certain topics. On some occasions, researchers need to find all documents relevant to a given query. For example, researchers who are new to "deep learning" may need to retrieve as many documents in their field as possible from scholarly databases, while social media monitors may want to retrieve all web pages containing selected topic words using a search engine. The data source type, including commercial search engines and scholarly databases, has three attributes: the only way to collect data from the data source is retrieving documents based on the relevance match between documents and given query; the number of results retrieved per query (denoted as *nl*) is limited; and the number of all documents matched (denoted as *nw*) is much bigger than *nl*. As *nw* is much bigger than *nl*, it is impossible to collect all the data from a limited accessible data source with only one query and the only clue for data collection is the given query. To solve this problem, query expansion techniques are used to expand the query set.

In this paper, the authors explore two key issues: How to solve the problem that the number of retrieved documents is limited by data source and how to generate a set of terms for query expansion, which will minimize the number of queries used for obtaining all the relevant documents from the limited accessible data source. Specifically, this paper proposes a query expansion method aiming to exhaustively collect data from a data source based on a minimum coincide tree and a method for selecting words based on LCA to solve these problems. In the field of information retrieval, there are many definitions of relevance. To facilitate the evaluation, the authors consider every document containing the query as a relevant document.

A segmentation method based on a minimum coincide tree

In terms of vector space, because of the proximity of spatial distance, the documents of the corpus tend to form many clusters. For the purpose of acquiring data from a limited accessible data source, the number of each retrieval result has an upper limit; therefore, reducing duplication of retrieval results is necessary to minimize the total number of queries for retrieving data. If the discrimination between clusters is large, then the overlap among clusters is low; that is, the queries used for obtaining all the documents are of a small number. Therefore, if there is a suitable method to divide an entire document set into many highly distinguished clusters, then the problems listed earlier can be solved.

The model assumes that there exists a document collection $C = D_1, \cdots, D_N$ where D_i is a document/data. Given query Q_{ori}, the task is to return all of the documents (denoted as C_r), which is relevant to the Q_{ori}. In this paper, the assumption is that the document which contains Q_{ori} is relevant to query Q_{ori}. Although this relevance is not perfect, it should work well on many occasions. Until the relevant documents are collected, C_r is unknown. With Q_{ori}, only a fixed number of documents can be collected. To identify more relevant documents, more queries for retrieval should be used; that is, the query set which is originally denoted as $QS = Q_{ori}$ should be expanded.

The proposed method resembles document filtering where the nodes represent words or phrases, each path represents a filter criteria and each document treats the original search term Term_1 as a starting point and moves down. If the document contains the vocabulary represented by the node, then it moves to the next term. In Figure 1, all the documents that contain the word Term_1 and Term_3 can be segmented into document Set 1 and then into Set 2 if they contain the terms Term_6 and Term_7. Documents containing

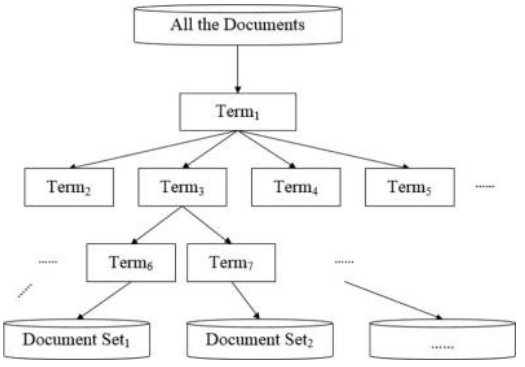

Figure 1.
The process of segmenting document collections by terms

the query terms $Term_1$ and $Term_3$ are gathered by obtaining documents from document Sets 1 and 2. Similarly, all the documents containing query terms $Term_2$, $Term_3$, $Term_4$ and $Term_5$ can be obtained. With this method, all documents containing the query term $Term_1$ are accessed. If the overlap between the document sets is low, then the number of queries used for obtaining all the documents containing $Term_1$ is small. Therefore, the researchers need to select appropriate words to construct a minimum overlap tree structure which makes the overlap between different document collections under the leaf nodes the lowest. These leaf nodes can be chosen as terms used for query expansion.

Query expansion method based on local context analysis

Query expansion methods could be grouped into global analysis-based methods, local analysis-based methods and query log analysis-based methods (Li *et al.*, 2010). With a limited accessible data source, researchers are unable to obtain the relationships among all the terms; thus, the global analysis method is not applicable and query log analysis-based methods are also not applicable in the current scenario. Therefore, LCA is used to obtain expansion terms for the segmentation of the document collection.

LCA is utilized to select query expansion terms. A similar work was performed by Xu and Croft (1996) for the purpose of improving information retrieval performance, which is different from the goal of the current paper. To collect relevant documents exhaustively, other criteria were developed to select query expansion terms. The current research tends not to use terms with a higher similarity with the initial query terms because similar terms would return highly overlapped relevant documents. The method developed in this paper is referred to as *inverse local context analysis* (ILCA). The following three steps are taken:

(1) Retrieve a list of top-ranked relevant documents S for query Q_{ori} in collection C.

(2) Rank the terms within S using *f(c, QS)* (see Formula 1) in reverse order. In Formula 1, $co\_degree(c, w_i)$ is the number of co-occurrences between term c and query terms w_i:

$$f(c, Q) = \prod_{w_i \in Q} (\lambda + co\_degree(c, w_i))^{idf(w_i)} \quad (1)$$

(3) Add the best k concepts to QS.

k can be set empirically according to the size of remaining documents to be gathered; that is, $k = f(|RD|)$, in which $|RD|$ is the size of the remaining documents to be gathered and k is negatively correlated to $|RD|$. In formula 1, co_degree(c,w) is different from the previous formula, the frequency of query term w in document d is replaced with the length of document d.

In summary, by combining ILCA and the segmenting method of document sets mentioned above, the procedure for obtaining data from a limited accessible data source is shown in Figure 2.

The method of retrieving all of the retrieved documents N from a limited accessible data source includes several steps:

- *Step 1*: Perform an initial retrieval on C to get the top-ranked set S for Q. Add the best k concepts $(w_1, w_2, \cdots, w_k)$ which are selected by the ILCA to Q and obtain k

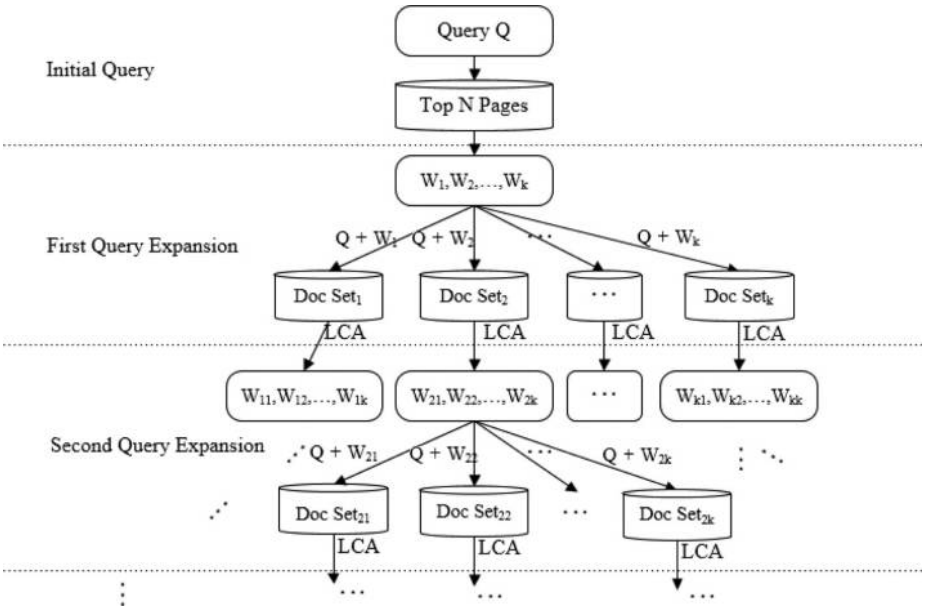

Figure 2. Methodology for extracting data from a limited accessible data source

document sets by performing k times query expansion using the query formula $Q + w_i$ for each query expansion.

- *Step 2*: If the total number of unique documents is less than N, then perform the query expansion by selecting m expansion terms from each document set of the k document sets above. If the total number of unique documents is equal to N, then the process finishes.
- *Step 3*: Repeat Step 2 until the total number of returned documents is equal to N.

In addition, Xu and Croft (1996) compared a series of user queries and found that users prefer to use nouns and noun phrases to express their information needs and that noun phrases are also more suitable for user interaction. Therefore, it is good to add a restriction when selecting expansion terms: The selected terms should be nouns or noun phrases such as "Washington" or "Stanford University".

Empirical evaluation

The method proposed in this paper aims to collect documents from a limited accessible data source. As discussed above, a limited accessible data source, including commercial search engines and online scholarly databases, has three attributes:

(1) the primary way to collect data from the data source is retrieving documents based on the relevance match between documents and a given query;

(2) the number of results retrieved per query (denoted as *nl*) is limited; and

(3) the number of all documents matched (denoted as *nw*) is much bigger than *nl*.

To evaluate the method, the research simulated the technological processes of a limited accessible data source. Both search engines and online scholarly databases are good choices, as they both meet the core attributes of a limited accessible data source, as defined earlier. In this paper, a local search engine was used to test the performance of the method. The authors did not use a scholarly database for the evaluation, as it is almost impossible to collect enough full-text documents to simulate a real online scholarly database.

To set up the experiment, the authors built a local search engine. The data set used in the experiment contains 1,410,000 news documents published on the Internet. Lucene was used for the indexing and searching. The authors implemented the core technological process of commercial search engines for this experiment.

Figure 3 demonstrates the experiment in the local search engine and the process of acquiring data from a limited accessible data source.

Data sets

In this experiment, the Sogou-news data set (Sogou, 2012) was used for the evaluation. The data set, released by Sogou Laboratory, contains 1,410,000 news documents published on the Internet. Each document contains several fields, including title, URL, content and news id.

To construct a local search engine for experimentation, the Chinese lexical analyzer ICTCLAS (http://ictclas.nlpir.org) was used for word segmentation and POS tagging. The stopwords list from HIT-SCIR was used to remove stopwords (StopWords List [EB/OL], 2013). Lucene 4.5 (http://lucene.apache.org) was used to index all of the documents in the Sogou-news data set. The fields used for indexing were the title and content. To simulate the effect of a commercial search engine, the maximum number of results per retrieval was limited to 700.

In all, 50 queries randomly selected from the top 2,000 most frequent noun phrases in our corpora are used for evaluating the performance. The queries included America, Shanghai and automobile.

Evaluation method

To achieve the purpose of this paper, the proposed method should exhaustively gather all of the retrieved documents from the limited accessible data source. Under this

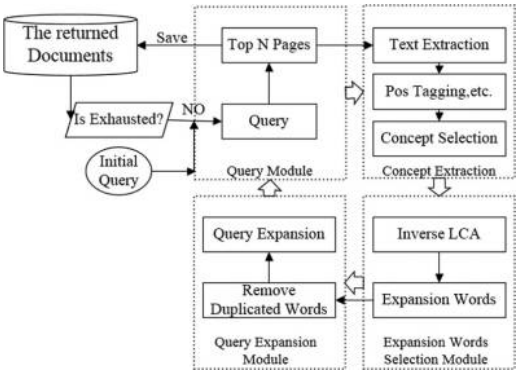

Figure 3.
Framework for exhaustive collection of data from a limited accessible data source

prerequisite, the goal is to minimize the number of queries used for obtaining all of the relevant documents. The evaluation criterion is that 95 per cent of the relevant documents are retrieved. The reason for the 95 per cent level is that after initial queries, it was decided that gathering 100 per cent of documents was impossible.

Baseline method

Three methods were used as the baseline methods for the query expansion:

(1) *Term frequency method* (TF) (Sparck Jones, 1972) selects the most frequent words for query expansion.

(2) *Term frequency method-inverse document frequency method (TF-IDF)* (Sparck Jones, 1972) calculates the weights for each word in the retrieved documents, words with high TF-IDF weights will be used for query expansion.

(3) *Term clustering method (TC)* clusters all words in retrieved documents using the K-means algorithm and the centroid of each cluster will be used for expansion. The distance between two terms is measured using co-occurrence frequency.

Setting of the number of expansion terms per round

To improve the recall of the retrieved results and reduce the overlap among the returned document sets, the expansion-based method is used to generate queries for retrieval. To implement the proposed method, the first problem is to determine how many terms should be selected to generate queries. The returned results from the initial query may cover multiple topics, despite the exact count not being known. It is assumed that a proper number of terms selected to generate new queries, according to the distribution of topics, will help improve the recall. The number of query expansion terms selected from the initial results has a major impact on the total number of documents users retrieve.

To test whether the hypothesis was established or not, three initial queries were used: "region/n", "economy/n" and "fund/n". The authors tested the method's performance using a different number (denoted as l) of expansion terms selected per round. After the nth round, there were $n \times l$ queries generated to retrieve documents. The documents obtained by the top 500 queries were compared. Figure 4 shows that the number of expansion words selected from the initial query results had little impact on the total number of documents returned by the top 500 queries. Figure 5 demonstrates that the number of expansion terms selected from the initial query results had low impact on the total number of queries needed for obtaining all of the retrieved documents. Thus, the assumption above does not hold. However, the number of expansion words selected from the initial query results had a high impact on the speed of accessing documents. A

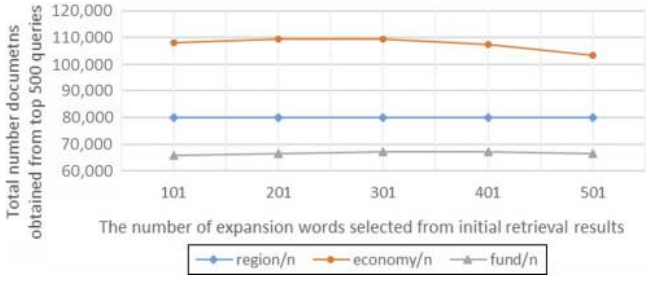

Figure 4. The relationships between the number of initial expansion words and top 500 retrieval results

proper number of expansion words selected in the initial round is beneficial to speed the gathering process.

In this experiment, the number of expansion terms selected for the first round was set to 500. The first round generated 500 queries. The overlap between each result set and the total result set obtained was calculated. If the overlap ratio was greater than 0.7, then no expansion term was selected from the document set. Documents retrieved after several rounds tended to focus on fewer topics than those retrieved in the first round; therefore, the number of expansion words selected after the second query expansion was set to 100.

Results and discussion

The experimental results of the four query expansion techniques: TF, TF-IDF, TC and ILCA are shown in Table I.

In Table I, the first column lists the queries used in the experiment. The second column provides the number of documents retrieved (denoted as $|WD|$) using the corresponding query in the first column. The third column shows $|WD| * 95$ *per cent*, which is the number of documents needed for acquiring. The other four columns give the sizes of queries needed for gathering 95 per cent of all documents found in the data source using the four query expansion methods. As described above, a smaller query size indicates better performance. Evaluation results obtained in the experiment show that ILCA uses a lower number of queries to gather the needed documents. Results show that LCA is 23.3 per cent more effective than the TF-based method, 35.3 per cent better than the TF-IDF based method and 25.6 per cent more efficient than the term clustering method.

As shown in Table I and Figure 6, the LCA method uses the least number of queries compared with other methods. For different initial queries, the performance of LCA is more stable than other methods. The standard deviation of query size needed with each method was calculated, and LCA resulted in a smaller standard deviation (124.13) than the TF-based method (153.01), the TF-IDF-based method (176.92) and the clustering-based method (143.30). Among all the methods, LCA offers the best performance. More queries are used by the clustering-based method than the TF-based method, but a two-tailed *t*-test shows that the difference is not significant. The TF-IDF-based method gives a significantly worse performance than other methods.

Term selection for expansion in the four methods was also analyzed. The TF-based method preferred common words which may cause significant overlap among results retrieved using these words. However, as these words are quite common, they performed well in the pilot rounds. The TF-IDF-based method used words which are not so common but have small coverage in the document set. The clustering-based method

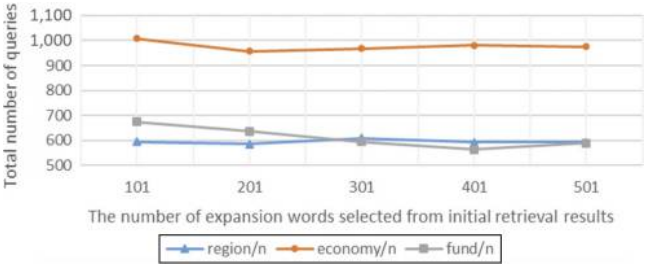

Figure 5.
The relationships between the number of initial expansion words and total number of queries

Query	Number of returned documents |WD|	|WD|*95%	Number of queries needed for exhausting 95% of the documents in a search engine TF	TF-IDF	TC	Inverse LCA
Medium/n	89,246	84,783	765	832	748	579
World/n	85,229	80,967	741	887	776	571
Staff/n	90,365	85,846	835	1,037	807	505
Society/n	96,394	91,574	852	970	775	662
America/ns	98,338	93,421	923	1,185	847	686
Government/n	96,935	92,088	834	912	777	621
Center/n	92,891	88,246	830	992	790	592
Shanghai/ns	72,882	69,238	732	871	712	552
Automobile/n	65,864	62,571	745	863	757	716
Department/n	98,394	93,474	833	1,014	806	600
Project/n	82,300	78,175	854	1,025	958	636
Environment/n	61,994	58,894	723	916	679	426
News/n	66,371	63,052	612	715	690	493
Country/n	133,058	126,405	1,151	1,291	1,137	794
Internationality/n	110,930	105,383	911	1,100	962	728
Economy/n	131,839	125,247	1,141	1,383	1,224	975
Region/n	86,618	82,287	907	1,101	747	592
Expert/n	61,824	58,733	572	691	570	451
Fund/n	71,906	68,310	694	794	749	589
Industry/n	67,382	64,013	626	717	623	603
			.....			
Average	*88,038*	*83,635.35*	657.94	780.04	678.18	504.76
Increase			*−23.28%**	*−35.29%**	*−25.57%**	

Note: *Indicates significant increase, two-tailed *t*-test with confidence = 0.05

Table I.
Number of queries needed for gathering 95% documents

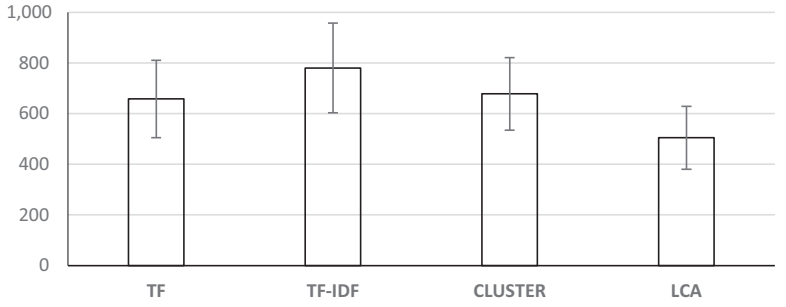

Figure 6.
Queries needed for gathering 95 per cent of relevant documents from a search engine

did not perform well. A more effective clustering algorithm may help to improve the performance, but this effort will be left for future work.

Conclusions

Limited accessible data sources, such as commercial search engines and online databases, have become important resources to create corpus and news reports data.

However, limited accessible data sources inhibit the number of results returned, which prevent users from obtaining all of the necessary documents.

To solve this problem, the researchers used a segmentation method of document sets based on the minimum coincide tree and utilized ILCA to select a low level of co-occurrence words with original query terms as expansion words to exhaust all of the relevant documents from a limited accessible data source. The experimental results show that the proposed method is able to gather most of the retrieved documents from a limited accessible data source with the least number of queries and is relatively insensitive to the original query terms. This method, thus, can improve the efficiency of obtaining data from a limited accessible data source.

While ILCA is proved to be useful in this work, it still has some shortcomings. Because of the high time cost of term calculation, compared to TF, DF and TF-IDF, ILCA requires more time for term selection. ILCA is also needed to be further improved for a better performance. In the future, the authors will further improve ILCA and explore additional methods for gathering data from a limited accessible data source.

References

Baroni, M. and Bernardini, S. (2004), "BootCaT: bootstrapping corpora and terms from the web", paper presented at Proceedings of the 4th Language Resources and Evaluation Conference (LREC), Lisbon.

Biancalana, C. and Micarelli, A. (2009), "Social tagging in query expansion: a new way for personalized web search", *International Conference on Computational Science and Engineering*, Vol. 4, pp. 1060-1065, doi: 10.1109/CSE.2009.492.

Cao, G., Nie, J.Y., Gao, J. and Robertson, S. (2008), "Selecting good expansion terms for pseudo-relevance feedback", *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, Singapore*, pp. 243-250.

Chum, O., Philbin, J., Sivic, J., Isard, M. and Zisserman, A. (2007), "Total recall: automatic query expansion with a generative feature model for object retrieval", *Computer Vision, ICCV 2007, IEEE 11th International Conference, Rio de Janeiro, Brazil*, pp. 1-8.

Crespo, A.M., Mata, V.J. and Maña, L.M. (2012), "Improving image retrieval effectiveness via query expansion using MeSH hierarchical structure", *Journal of the American Medical Informatics Association*, Vol. 20 No. 6, pp. 1014-1020.

Datta, R., Joshi, D., Li, J. and Wang, J.Z. (2008), "Image retrieval: ideas, influences, and trends of the new age", *ACM Computing Surveys (CSUR)*, Vol. 40 No. 2, p. 5.

Dou, W., Wang, K., Ribarsky, W. and Zhou, M. (2012), "Event detection in social media data", *IEEE Visweek Workshop on Interactive Visual Text Analytics-task Driven Analytics of Social Media Content, Washington, USA*, pp. 971-980.

Feng, P. and Huang, M.H. (2011), "Query expansion of pseudo relevance feedback based on feature term extraction and correlation fusion", *New Technology of Library and Information Service*, Vol. 27 No. 1, pp. 52-56.

Grefenstette, G. (1999), "The World Wide Web as a resource for example-based machine translation tasks", *Proceedings of the ASLIB Conference on Translating and the Computer*, Vol. 21, available at: www.mt-archive.info/Aslib-1999-Grefenstette.pdf (accessed 26 February 2016).

Hu, J., Wang, G., Lochovsky, F., Sun, J.T. and Chen, Z. (2009), "Understanding user's query intent with Wikipedia", *Proceedings of the 18th International Conference on World Wide Web, ACM, New York, NY*, pp. 471-480.

Huang, C.K., Chien, L.F. and Oyang, Y.J. (2001), "Interactive web multimedia search using query-session-based query expansion", *Advances in Multimedia Information Processing*, Springer, Berlin, pp. 614-621.

Huang, X., Huang, Y.R., Wen, M., An, A., Liu, Y. and Poon, J. (2006), "Applying data mining to pseudo-relevance feedback for high performance text retrieval", *International Conference on Data Mining, Hong Kong, HK*, pp. 295-306.

Jing, Y. and Croft, W.B. (1994), "An association thesaurus for information retrieval", *Proceedings of RIAO, Rockefeller University, New York, NY*, Vol. 94 No. 1994, pp. 146-160.

Kilgarriff, A. and Grefenstette, G. (2003), "Introduction to the special issue on the web as corpus", *Computational Linguistics*, Vol. 29 No. 3, pp. 333-347.

Ku, L.W., Liang, Y.T. and Chen, H.H. (2006), "Opinion extraction, summarization and tracking in news and blog corpora", paper presented at Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs, California.

Lee, K.S., Croft, W.B. and Allan, J. (2008), "A cluster-based resampling method for pseudo-relevance feedback", *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 235-242.

Li, Y.N., Wang, B. and Li, J.T. (2010), "Query expansion in search engines: a survey", *Journal of Chinese Information Processing*, Vol. 24 No. 6, pp. 75-84.

Liu, V. and Curran, J.R. (2006), "Web text corpus for natural language processing", paper presented at the Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy.

Parapar, J. and Barreiro, Á. (2011), "A cluster based pseudo feedback technique which exploits good and bad clusters", *Advances in Artificial Intelligence – 14th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2011*, La Laguna, 7-11 November, pp. 403-412.

Qi, X. and Davison, B.D. (2009), "Web page classification: features and algorithms", *ACMComputing Surveys (CSUR)*, Vol. 41 No. 2, p. 12.

Qiu, Y. and Frei, H.P. (1993), "Concept based query expansion", *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA*, pp. 160-169.

Robert, L. (2009), "The web as corpus versus traditional corpora: their relative utility for linguists and language learners", in Baker, P. (Ed.), *Contemporary Corpus Linguistics*, Continuum International Publishing Group, New York, NY, pp. 289-300.

Sharoff, S. (2006), "Creating general-purpose corpora using automated search engine queries", in Baroni, M. and Bernardini, S. (Eds), *WaCky*, Bologna, Italy, pp. 63-98.

Sogou (2012), *Sogou Web Corpus*, available at: www.sogou.com/labs/dl/q-e.html (accessed 17 September 2013).

Sparck Jones, K. (1972), "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, Vol. 28 No. 1, pp. 11-21.

Sparck Jones, K. and Barber, E.O. (1971), "What makes an automatic keyword classification effective?", *Journal of the American Society for Information Science*, Vol. 22 No. 3, pp. 166-175.

StopWords List [EB/OL] (2013), *HIT-SCIR StopWords List [EB/OL]*, available at: http://ir.hit.edu.cn (accessed 28 September 2013).

Tang, X.B. and Fang, X.K. (2014), "Microblog retrieval based on semantic query expansion", *Information and Documentation Services*, Vol. 35 No. 2, pp. 34-38.

Wang, C. (2011), "The hotspot detection and topic tracking based on public opinion analysis system", Master's thesis, University of Electronic Science and Technology of China, China.

Wang, C., Zhang, L. and Zhang, H.J. (2008), "Learning to reduce the semantic gap in web image retrieval and annotation", *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, NY, pp. 355-362.

Wilson, T.D. (2000), "Recent trends in user studies: action research and qualitative methods", *Information Research*, Vol. 5 No. 3, pp. 5-26.

Xu, J. and Croft, W.B. (1996), "Query expansion using local and global document analysis", *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4-11.

Zhu, M. and Xie, S.H. (2003), "Translating Chinese words and phrases into English with the help from search engine", *Shanghai Journal of Translators for Science and Technology*, Vol. 1, pp. 59-62.

Zhu, Y.J. (2012), "The construction of public opinion analysis system based search engine", Master's thesis, University of Electronic Science and Technology of China, China.

About the authors

Wei Lu is currently a Professor and Vice Dean of the School of Information Management at Wuhan University. He holds a PhD degree from Wuhan University. He has published dozens of papers in journals and conferenced, including the *Journal of Information Science*, *Aslib Proceedings* and *SIGIR*. His research interests include information retrieval, digital library and knowledge management.

Xinghu Yue is currently a Graduate Student in the School of Information Management at Wuhan University. His research interests involve information retrieval and Java programming.

Qikai Cheng is a Doctoral Candidate in the School of Information Management at Wuhan University. His research interests include information retrieval, digital library and natural language processing. Qikai Cheng is the corresponding author and can be contacted at: chengqikai0806@163.com

Rui Meng is currently a Graduate Student at Wuhan University. His research interests include information retrieval and digital libraries.