ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

Contextual Query Expansion for Acquiring Web Documents

Bidisha Roy[1], Priyanka Upadhyay[2]

Associate Professor, St.Francis Institute of Technology, Mount Poinsar, S.V.P. Road,Borivali (w), Mumbai, India[1]

Student, M.E.(Computer Engineering) St. Francis Institute of Technology, Mount Poinsar,

S.V.P. Road,Borivali (w),Mumbai, India[2]

Abstract: Query expansion is an information retrieval technique in which new query terms are added to the original query terms to improve search performance. Contextual query expansion is major issue in today's era. In this paper, contextualization is achieved by performing document extraction and terms extraction activities to the particular domain information source. User query is expanded using document extraction and terms extraction activities. Document extraction is achieved by BM25 retrieval function. It ranks set of documents based on query terms appearing in each documents .Now second function is terms extraction process, in this, terms in the top returned documents are weighted using sub linear terms frequency scaling formula which is used to weight the terms in the expanded query derived from original query which will cope with the term mismatch problem in specific domain. Hence, this paper presents a proposal to make web searches adaptive to the context of the user's query, thus improving query results. The proposed approach makes the context acquisition faster, In addition the results of the query engine with and without the contextual information showed improvements in the precision and search length of the web results.

Keywords: BM25 ranking model, Context, Information retrieval, Internet, Query Expansion, Sub linear term frequency scaling function.

I. INTRODUCTION

When searching for information on internet, people usually do not obtain much relevant information in the initial search and they need to modify queries and search again until they get satisfactory results. Query expansion is an automatic search technique that extracts useful terms from selected documents to improve the search result.
In now a days, people are relying more and more on the web for their diverse needs of information. Internet users usually describe their information needs by a few keywords in their queries, which are likely to be different from those index terms of the documents on the web. This problem is general in Information Retrieval (IR) systems and has been documented before the popularization of the web. New or intermittent users often use the wrong words and fail to get the actions or information they want [1].As a consequences, in many cases documents returned by search engines are not relevant to the user information need. This raises a fundamental problem of term mismatch in information retrieval, which is also one of the key factors that affect the precision of the search engines.

The search engines like google and yahoo are so famous that they are in use now and then for searching various type of information available on web. A web has become largest available data set in public domain to the extent that now –a-days; all are using a term "Information Explosion "as the data indexed by the search engines is so huge. Information retrieval (IR) is a scientific research field concerned with the design of models and techniques for selecting relevant information in response to user queries within a collection of documents [2].

The specification of user information need is completely based on words in the original query in order to retrieve documents having these words. Suchapproaches have been limited due to the absence of relevant keywords as well as the term variation in documents and users query. These issues have been addressed in semantic IR approaches which take into account the meaning of terms and semantic relatedness between senses in termino-ontological resources for enhancing the document/query representations or users query expansion. This paper describes ,how to extracts the document for indexing i.e. BM25 term weighting model [3].BM25 is retrieval function that ranks a set of document based on the query terms appearing in each document which results in query expansion. This paper presents an approach for getting additional terms to the user context need for providing a more precise description of users information need.

II. LITERATURE SURVEY

Different ways proposed by different authors for contextual search on Internet using query expansion. In [4], authors introduce a novel approach to search for biomedical information contextual search using contextual semantic information. More specifically, authors propose to combine the contextual semantic information in documents and user queries in an attempt to improve the performance of biomedical information retrieval (IR) systems. In this author presents context sensitive information retrieval(IR)approach, terms describing concepts are extracted from each document using several biomedical terminologies .Preferred terms denoting concepts are used to enrich the semantics of the document

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

content. The user query is expanded using terms extracted from the top ranked expanded documents via blind feedback query expansion method. This blind feedback query expansion is achieved by BM25 retrieval function and after that DFR (Divergence from randomness) term weighting model is used to expand the terms in the expanded query derived from original query.

In [5], author introduces a novel automated query expansion mechanism. In this, query expansion process, the semantics relations between the original query and expanding words is developed, in the context of utilized corpus. This query expansion process utilizes LSA (latent semantic analysis) for more efficient and reliable query expansion process.In LSA (latent semantic analysis), each meaning of words or a sentences is modeled as a vector in the semantic space, where the meaning of sentences is the sum of its words vectors. Here, SVD (Singular value decomposition) Matrix is constructed where rows represent the unique words and Column represent the paragraph. Words are then compared by taking the cosine of the angle between the two vectors formed by any two rows. Here, Values near to 1 represent very similar words while values near to 0 represent very dissimilar words. In this paper, for the purpose of evaluating system a corpus composed of articles from the Agency France Press (AFP) Arabic Newswire is used.

In [6], author propose a new method for query expansion based on user interactions recorded in user logs. The idea is to extract correlations between query terms and document terms by analysing user logs. These correlations are then used to select high-quality expansion terms for new queries.

The idea behind implementation of this paper is based on context sensitive information retrieval approach for query expansion. To achieve contextual query expansion for acquiring web documents, this paper describe two approach first one is document extraction and other is terms extraction. For document extraction, each document in the collection is analysed to extract most relevant document to the user query. To achieve document extraction, it perform BM25 term weighting model , which will be useful for selecting most related document from information source which is specific to the domain to measure the degree of description of each concept to the semantics of the document which result in best user query searching in the document. Now, we perform data pre-processing activities (stop word removal, tokenization) on top returned document by document extraction process which are most related to the user queries [7]. Now, user perform second function which is terms extraction. Terms extraction is an activity of obtaining useful terms for query expansion from top returned documents of document extraction activities. Here, Sub linear terms frequency scaling formula is used to weight the terms in the expanded query derived from original query.

For search to become sensitive to a domain context, the strategy proposed in this work is based on following hypothesis: “the terms most often found in an information resources that is representative of a domain are more likely to also be present in other related and relevant documents available in Internet. Therefore, when using these terms to expand queries made by users, it is possible to obtain more useful search results”. But using that each resources which represent the context can have different topics in its content, a simple extraction of terms based on their occurrences can result in combination of different terms of different subjects in a query, possibly reducing probability of obtaining useful results in the search. To minimize this risk, it is proposed to perform document extraction using BM25 term weighting model. Document extraction will extract documents which are related to the query given by user from entire knowledgebase source. It can be achieved by BM25 ranking function that ranks a set of documents based on the query terms appearing in each document [8].

III. PROBLEM DEFINATION

Information retrieval is a scientific research field concerned with the design of models and techniques for selecting relevant information in response to user queries within a collection (corpus) of documents. So, In order to make search results more relevant considering the users' behavior when building a query, it is used an information retrieval technique known as query expansion.
In this Project, terms are added in the original query made by the user in an attempt to provide a greater contextualization, and retrieving more useful documents.

IV. PROPOSED SYSTEM

The existing architecture (showed in Fig. 1) is organized in three modules: Knowledge Base Configuration, Information Extraction and Search. In Knowledge base Configuration module, the learning domain context is obtained through the use of the contextual sources, which are the files published as educational resources (articles, book chapters, lecture notes, publications in general) or the messages exchanged among the participants of the learning activities (messages obtained from the use of communication tools). Next Information extraction module, the objective of this module is to identify the main terms of the contextual information obtained from the Knowledge Base Configuration module, and to provide a list of (additional) terms for the search module. Finally, the search module receives the keywords expanded using the terms extracted in the Information Extraction module, and the resulting query is executed in the web search engine [8].

In the proposed system, we have replaced Information Extraction (Segmentation and Clustering) activities with document extraction (BM25 Ranking Module) and data pre-processing activities. Hence, proposed system is divided in to five modules: Knowledge Base Configuration, Document extraction, data pre-processing, terms extraction and query expansion.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

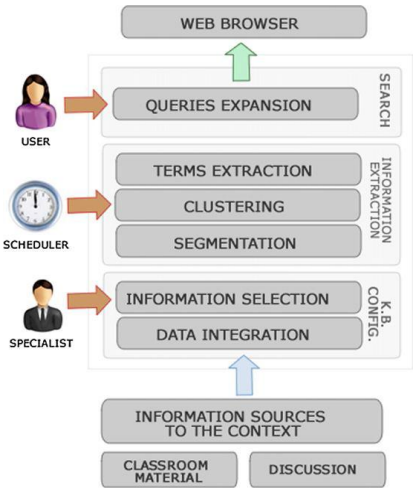


Figure 1. Existing Architecture [8]

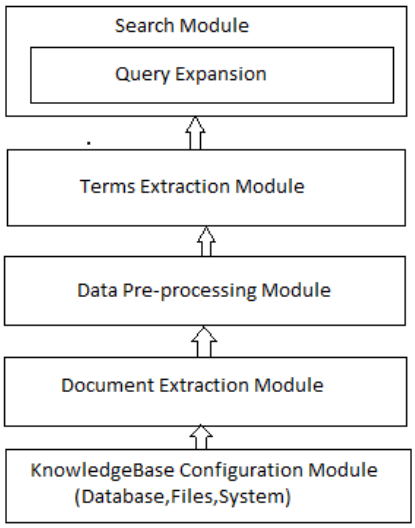


Figure 2. Proposed architecture

1) Knowledge Base Configuration Module: The domain context is modelled with the use of existing resources such as databases, miscellaneous files (articles, book, chapters, and publications in general) or information systems (including data integrations performed through web services). Any information source that contains textual content and information, which represent the specific domain, can be used for this purpose. Here, software engineering domain is considered.

2) Document Extraction Module: This will be achieved by BM25 ranking function. It is a ranking function, which ranks given documents according to their scores for a given query by user. Score of a document D will be calculated as:

Given a query Q containing keywords $q_1, q_2 \ldots q_n$, the BM25 Score of a document D is:

$$\text{Score}(D, Q) = \sum_{I=1}^{n} IDF q_i \cdot \frac{f(q_{i,D}).(k_1+1)}{f(q_i,D)+k_1.(1-b+b.\frac{|D|}{avgdl})} \quad (1)$$

Where, f (q_i, D) is q_i's term frequency in the Document D.|D| length of document D in words and avgdl is the Average document length in text collection From which Documents are drawn.k_1, b are free Parameters, Usually Chosen in advanced optimization $Ask_1 \in$ (1.2, 2.0) and b=0.75.IDF (q_i) is the IDF (Inverse document Frequency) Weight of the query $Termq_i$.It is usually computed as:

$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}, \quad (2)$$

Where, N is total number of documents in the collection (q_i) is the number of Documents containingq_i.

3) Data Pre-processing Module: It is necessary to apply some activities of text pre-processing before the extraction of terms : (1)the tokenization ,the process of breaking a stream of text into words,phrases,symbols and other meaningful elements called tokens,(2)the removal of stop words ,a list of common or general that have little value in the text and must not be extracted and(3) stemming, the process for reducing inflected (or sometimes derived)words to their stem, base or root form [7].

4) Terms Extraction Module: The extraction of terms of context is done by calculating the weights of the terms and extracting the n terms with highest weight, where n is the maximum number of terms that can be used in the expansion of query. The weight calculation is performed with the formula of sub linear term frequency scaling (3) [7].

$$wf_{t,d} = 1 + \log tf_{t,d}, \text{ if } tf_{t,d} > 0 \quad (3)$$

Where, $tf_{t,d}$ is frequency of term t in document d and $wf_{t,d}$ is weight of term t in document d.

5) Search Module: The search modules receive keywords to perform search on the web. The original query is expanded using terms extracted in the information extraction module, and the resulting query is executed in web search engine.

The original query can be expanded in two ways. The first is the automatic expansion, in which the original query is expanded n times. Each expanded query is executed in web browser and results are presented to the user in tabs.

The second mode of query expansion is the suggestion of terms, in which all extracted terms are presented as a suggestion. The user has to select terms of his/her interest that will be incorporated to the original query, performing the query expansion. When selecting the terms, the user can combine general terms with specific terms of the subjects.

V. RESULT ANALYSIS

For the purpose of evaluating our new system, we used the software engineering domain, a knowledgebase composed of pdf, books, research articles, lecture notes, pptfiles, of software engineering domain. The results were analysed according to two metrics (Precision and Search length) using existing architecture and proposed architecture [9]. Precision is a metric widely used in information retrieval and represents the fraction of relevant documents among all retrieved documents.

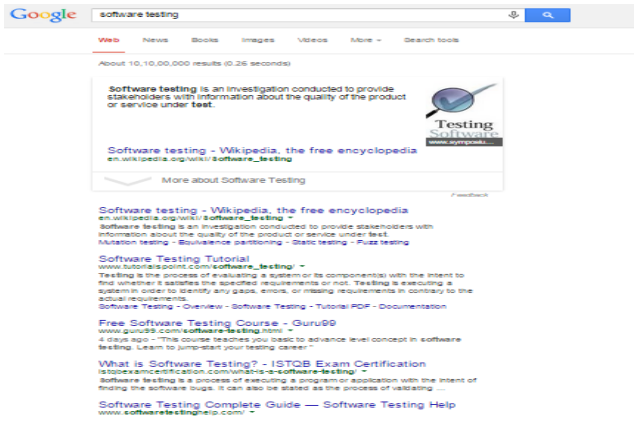
The full precision metric tries to consider the total amount of relevant information found in the first 10 results. The second metric was the search length, which reflects the number of non-relevant documents that the user must evaluate until finding a certain number of consecutive documents that are considered relevant. Therefore, lower

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

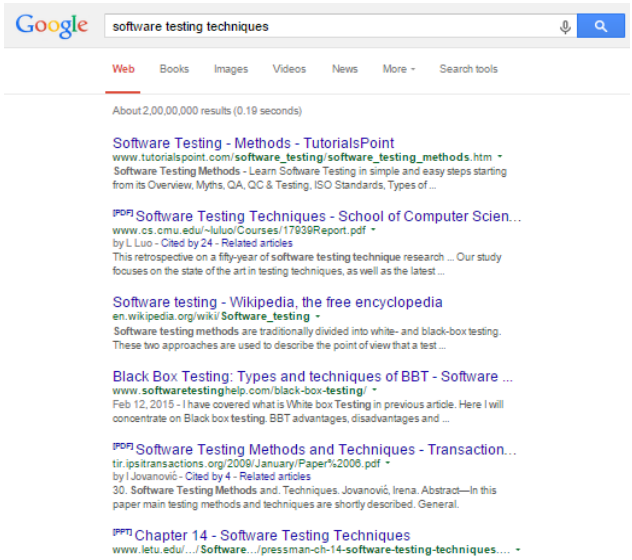
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

values correspond to less effort for the user to find relevant results.Here in this paper; analysis is done based on domain by users before and after expansion of the word using two techniques. For e.g. User gives 'Software testing' as a search query on the web search engines. On the submission of this search query over the web search engines user will get some web results. After using this Contextual web search tool user search query is expanded in 'Software testing techniques'. So, after this query expansion user will get some more relevant web documents than original search query hence, results over the query expansion shows better results in precision and search length of web documents.

Original search query: "Software testing"



Expanded search query: "Software testing techniques"



Experimental analysis is carried out with two techniques over original search queries. Analysis is carried out with Under 15 students with 10 search queries related to software engineering domain. Results shows improvements over average precision and search length metrics using both the techniques. Figure 3 and Figure 4below shows average improvements of precision and search length parameters respectively of original query search using both the techniques.

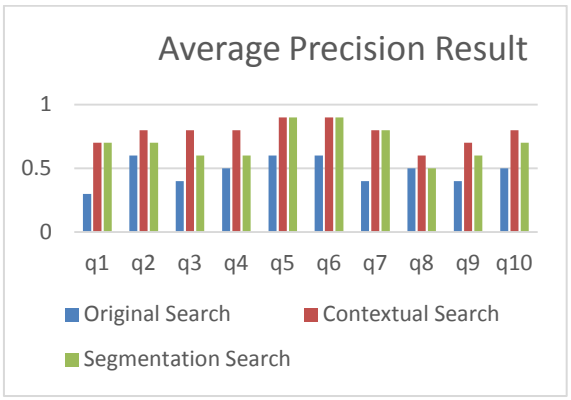


Figure 3: Average Precision Result

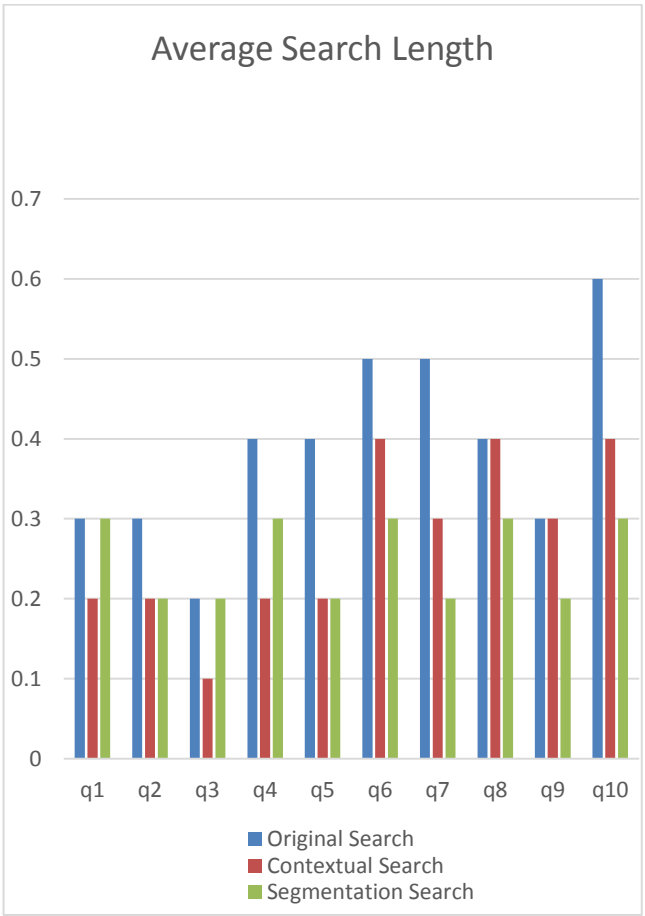


Figure 4: Average Search length Result

Comparing the Existing architecture and proposed architecture, we got that existing architecture took more time for processing as well as ,in the existing architecture we have to implement three techniques (segmentation ,clustering, terms extraction) where as in proposed architecture we have to implement two techniques (BM25 Ranking function, terms extraction)hence, existing architecture is less efficient than proposed architecture.

DOI 10.17148/IJARCCE.2015.4693

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

VI. CONCLUSION

The work presented in this paper, proposed a strategy for contextual search on internet using query expansion. To make query results more useful terms extraction, document extraction and query expansion techniques were considered.

The use of terms extraction, document extraction and query expansion activities provided more contextualized search results, increasing usefulness for users, helping them search for educational resources on the web. The contextualization is achieved through expansion of queries entered by users, adding in these queries terms extracted from knowledgebase configuration module, which is specific to the software engineering domain.

ACKNOWLEDGMENT

The authors would like to thank various authors whose papers have helped us to develop new idea and create the proposed architecture. We would like to thank our institutions for encouraging us to write the paper and create the proposed architecture.

REFERENCES

[1] Hang Cui,,Ji-Rong, Wen,Jian-Yun Nie ,,Wei-Ying Ma, " Query Expansion by Mining User Logs", IEEE Transaction On Knowledge and Data Engineering, vol15,no. 4, July/August2003.

[2] Renuka Nagpure et al, Int.J. Computer Technology & Applications, Vol 5 (4),1485-1490.

[3] S.E. Robertson, S. Walker, Hancock-M. Beaulieu, Okapi at TREC-7: AutomaticAd Hoc, Filtering, VLC and Interactive, In: Proceedings of Text Retrieval Conference, 1998, pp. 199–210.

[4] DuyDinh,LyndaTamine,"Towards a context sensitive Approach to searching information based on domain specific Knowledgesources",Web Semantics Science,Services and Agents on the World Wide Web 12-13(2012).

[5] Ahmed Abdelali,JimCowie,Hamdy S. Soliman,"Improving Query precision using semantic expansion",In formation Processing and Management 43 2007) 705-716.

[6] Hang Cui,Ji-RongWen,Jian-Yun Nie,and Wei-Ying Ma,"Query Expansion by Mining User Logs",IEEE transactions on Knowledge and Data Engineering, vol 15, No. 4, July/August 2003.

[7] Manning, C.D.,Raghvan ,P.,Schutez,H., "Introduction to Information Retrieval",Cambridge University Press,2008.

[8] João C. Prates,Sean W. M. Siqueira," Contextual Query Based on Segmentation and Clustering of Selected Documents for Acquiring Web Documents for Supporting Management", Association for Information Systems AIS Electronic Library (AISeL), May 2011.

[9] Tang, M. C., & Sun, Y. (2003). Evaluation of web-based Search engines using usereffortmeasures. LIBRES Research Electronic Journal, 13(2). <http://libres.curtin.edu.au/libres13n2/tang.htm>.