International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 03, March-2015

Context Sensitive Approach for Information Retrieval on Internet

Priyanka Upadhyay
Department of Computer Engineering
St. Francis Institute of Technology
Mumbai, India.

Bidisha Roy
Department of Computer Engineering
St. Francis Institute of Technology
Mumbai, India.

***Abstract*— Search engines are powerful tools to find information on the web. However, they commonly return many irrelevant documents when the user's queries are not specific enough. Web search engines have a key role in the discovery of information retrieval, but this kind of search usually performed using keywords and results do not consider the context. In the web environment, where collection tends to be enormous, it is so important to have robust queries. Query Expansion is one of the most important mechanisms in the information retrieval field. A typical short Internet query goes through a process of refinement to improve its retrieval power. Most of the existing QE techniques suffer from retrieval performance degradation due to imprecise choice of query's additive terms in the QE process. This paper describes the use of information extraction techniques applied in previously defined resources in order to expand the queries made by users and run these expanded queries in web search engines, getting more useful search results considering the domain context of the required information.**

Keywords*— *BM25 ranking model, Context, Information* Retrieval, *Query Expansion.

I. INTRODUCTION

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Many universities and public libraries are use IR system to provide access to books, journals and other public libraries are use IR system to provide access to books, journals and other documents. Web search engines are most visible IR applications. In search tools, the user usually reports an information need by entering keywords in search expressions. Although this, approach makes it easy for users to express their information need, but it also results in context less results. 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 technique, terms are added in the original query made by users in an attempt to provide a greater contextualization and retrieving more useful documents [1]. The terms that are added in the original query made by user, taken from the context of the information need can minimize the following query expansion issues: (a)the user needs to know the terms contained in the web pages/documents in order to get the results that will meet his/her information needs, and (b) words can have different meanings and the disambiguation of these concepts is from the responsibility of the user himself/herself ,that must modify the query by adding or changing the keywords that were used.

Another important issue that maximizes these problems is the user's behavior when using Internet search tools. Users browse a few results pages, often limited to the first 5 results [2], but statistics show that search expressions typically consist of few terms.65% of searches on the Internet contain from one to three terms [3]

In this paper, context is considered as any piece of information that can be used to characterize situation of entity. An entity is a person, a place or object that is considered relevant to the interaction between a user and an application needs, thus improving query results [4].

The contextualization is provided through the expansion of queries entered by users, adding to these queries some terms extracted from selected documents that are representative of the context of the information need.

II. RELATED WORK

Different ways are proposed by different authors for contextual query expansion for acquiring web documents.

In[5] , authors describe the use of information extraction techniques(segmentation and clustering) applied in previously defined resources in order to expand the queries made by users and run these expanded queries in web search engines, getting more useful search results, considering the domain context of the required information. The use of information extraction activities in existing resources in databases, archives and information systems can be considered in order to make search results more contextualized and therefore more useful to users.

The idea behind implementation of this paper is based on context sensitive information retrieval approach for query expansion.

In the existing system, information extraction activities are applied in entire knowledge base configuration module but in our proposed system, all these activities are performed on top ranked documents, which are related to the query given by user. To achieve this, BM25 term weighting model is used which is used for query expansion to measure the degree of description of each concept to the semantics of the documents which result in best query searching in the document.

Related work using query expansion with grouping approach is presented[6].These studies assume that, considering groups of similar document, documents in the same group are relevant to the same requeste.By classifying the query terms in one or more groups, the terms of these groups could be used to expand the query. The main criticism of this method is the poor results obtained when a small collection of documents is used for the groups or when the differences of vocabulary between relevant and not relevant documents are insufficient. Another problem is that a document can only belong to one group.

In [7], authors applied clustering techniques in the first X documents returned by original query, from which terms were generated for each group and displayed to the user to select the most appropriate set of terms for expansion.

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 information resources that are 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".

However, considering that each resources that represents a context can have different topics in its content, a simple extraction of terms based on their occurrences in the resources can result in a combination of terms of different subjects in a query, possibly reducing the probability of obtaining useful results in the search.

III. PROPOSED WORK

Proposed system is divided in to four modules: Knowledge Base Configuration, BM25 Ranking Function, Information Extraction and search.

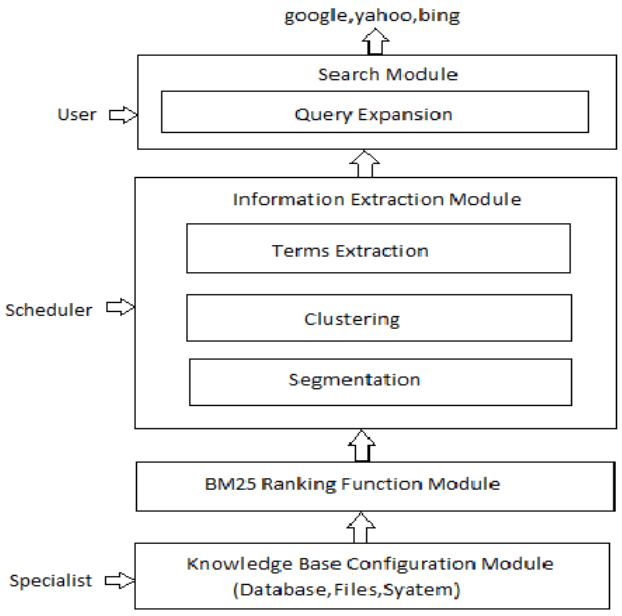

Figure 1. Proposed architecture

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, connectors, etc.). Any information source that contains textual content and information, which represent the domain context, can be used for this purpose.
2. *BM25 Ranking Function Module:* 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:

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

   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 as $k_1 \in$ (1.2, 2.0) and b=0.75.IDF (q_i) is the IDF (inverse document Frequency) Weight of the query term q_i.It is usually Computed as:
   IDF (q_i) $= \log\frac{N-n(q_i)+0.5}{n(q_i)+0.5}$, where N is total number of Documents in the collection (q_i) is the number of Documents containing q_i.
3. *Information Extraction Module:*

The objective of information extraction module is to identify main terms of all the contextual information obtained from knowledge base configuration module and to provide list of terms to search module. Two extractions of terms are executed, one for most frequently used terms in the context(Context extraction).Another for specific terms of each subject identified in context(subject extraction).

In both the situation it is necessary to apply data pre-processing activities before extraction of terms such as (a) Tokenization, the process of breaking stream of text into words, phrases, symbols and other meaningful elements called tokens. (b)The removal of stop words, a list of common or general terms that have little value in the text and must not be extracted. (C) Stemming, the process for reducing inflected or derived words to their stem base or root form .

The extraction of general terms of context is done by calculating the weights of terms and extracting n terms with highest weight, where n is 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 (2)[8] .

$$wf_{t,d} = 1 + log\, tf_{t,d}, \text{ if } tf_{t,d} > 0 \qquad (2)$$

0, otherwise.

Where, $tf_{t,d}$ is frequency of term t in document d and $wf_{t,d}$ is weight of term t in document d. The extraction of specific terms of each subject that is identified in the context requires the application of more information extraction activities. The first step is the routine of text

segmentation to divide the contextual information into sentences. The objective in applying the segmentation is to ensure that a particular textual content contains only one subject, thus ensuring that there are no terms that deal with different subjects in such content.Here,linear text tiling method is used[9] .In order to group these subjects it was used clustering. Clustering is a statically technique that allows automatic generation of grouping of data (documents).Here, document clustering is used, which is based on K-means clustering.

After all the contextual information grouped by subjects, the weights of all terms are calculated, considering each subject that was identified as a collection of independent documents, also using the weight expression(2).After the calculation of weights for all the terms ,the n terms with highest weight are extracted for each identified subjects.

4. 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, where n is equal to one (context expansion) plus the number of subjects that were identified in the selected context (subject expansion).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 (generated by the context extraction and subject extractions) 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.

IV. CONCLUSION

The use of information extraction and query expansion activities provided more contextualized search results, increasing usefulness for users, helping them search for educational resources on the web. This report focused on domain based query expansion. The work described in this proposed a strategy contextual query based on segmentation and clustering using BM25 retrieval function.

ACKNOWLEDGMENT

The authors would like to thank various authors whose paper has 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. Yates, R.B., Neto, B.R. (1999) Modern Information Retrieval , Addison Wesley, la ed.
2. Spink, A., Jansen, B.J. (2004) A study of web search trends. Webology, 1(2), Article 4. Available at:http://www.webology.ir/2004/v 1 n2/a4.html.
3. Experian Hitwise Searches ststistics (2010) http://www.hitwise.com/us/press/-releases/google-searches-feb-10/.
4. Dey, A. K., Abowd, G.D. (1999) Towards a better understanding of context and contextawareness.International symposium on Handheld and Ubiquitious Computing ,pp.304-307.
5. 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.
6. Bhogal, J., Macfatlane,A., and Smith,P.(20007).A review of ontology based query expansion.Information Processing and Management:An International Journal ,43(4),866-886.
7. Kang, J.W., Kang,H.K.,Ko,M.C.,Jeon,H.S.,and Nam,J.(2010).A term cluster query expansion Model based on classification information in Natural language information retrieval .In preceding of national conference on artificial intelligence and computational intelligence (pp.172-176).
8. Manning ,C.D.,Raghvan ,P.,Schutez,H.(2008) Introduction to Information Retrieval ,Cambridge University Press.
9. Hearst,M.A.(1997)TextTiling:Segmenting text into multi-paragraph subtopic passages.Computational Linguistics,pp 33-64.