Contextual information search based on domain using Query Expansion

Renuka Nagpure[#1], Reena Mahe[*2], Sumita Chandak[#3]
[#]*Dept. of Information Technology- Dept. of Information Technology,*
[#] *Atharva College of Engineering- Atharva College of Engineering,*
[#]*Malad(W), Mumbai, India- Malad(W), Mumbai, India.*
[1]renukanagpure@atharvacoe.ac.in
[3]sumitachandak@atharvacoe.ac.in
[*]*Dept. of Information Technology*
Atharva College of Engineering,
Malad(W), Mumbai, India.
[2]reenamahe@atharvacoe.ac.in

Abstract— **The Internet is one of the main information sources nowadays and information search is an important area in which many advances have been registered. As the context and semantics of the information in the web pages indexed depends on multiple factor, semantic search has become a complex task. One approach to improve web search results is to consider contextual information.The design of this system is knowledge source domain based web search, in this context sensitive IR approach terms denoting concepts are extracted from each document using several domain based terminologies. Preferred terms denoting concepts are used to enrich the semantics of the document via document expansion. The user query is expanded using terms extracted from the documents**

Keywords—Query Expansion, Information Extraction, Document Fusion, Indexing

I. INTRODUCTION

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 a 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 (corpus) of documents. Two main steps characterizing an IR process are document indexing and document–query matching [1]. The objective of the indexing stage is to assign to each document in the collection the set of words, terms or concepts expressing the topics or subject matter addressed in the document. The matching stage aims at identifying the most valuable documents that better fit the query.

II. EASE OF USE

The document–query matching between keywords from the user's query and documents is realized under the basic term independence assumption. The specification of the user information need is completely based on words figuring in the original query in order to retrieve documents containing those words. Such approaches have been limited due to the absence of relevant keywords as well as the term variation in documents and user's query (e.g. acronyms, homonyms, synonyms, etc.). 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 user's query expansion. In this paper the document is extracted for indexing by fusing model i.e.BM25 term weighting model [7]. BM25 is a retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document.

III. LITERATURE SURVEY

Recent advances in contextual query based on segmentation and clustering of selected documents for acquiring web documents for supporting knowledge management [3]. A survey of information for implementing contextual information retrieval shows main terms of the contextual information obtained from the knowledge base configuration module to provide a list of additional terms for the search module. Two extractions of terms are executed, one for the most frequently used terms in the context and other for specific terms of each subject identified in context [1]. Also, the manually assigned keywords applied for query expansion. There was no indication that are manually assigned keywords aided the users for query expansion or for imparting information about the document collection [6].

The idea behind this paper is based on context sensitive information retrieval approach for query expansion. During the indexing stage, each document in the collection is analyzed to extract the most significant concepts using several terminologies. The assumption behind multi-terminology based concept extraction is that the more concepts are found in several terminologies, the more they are important in the description of the document since they are well recognized in several sub domains of knowledge sources. For concept extraction, it adopt MaxMatcher, which is an approximate lookup based on dictionary matching [9]. Given a document, MaxMatcher will extract a set of terms or phrases denoting domain concepts as well as their corresponding concept unique identifiers (CUIs). However, MaxMatcher does not measure the importance of each concept for describing the semantics of the document. To achieve this, BM25 term weighting model [7] is uses which is used for query expansion 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.

In fusion, ranked lists are combined together by various means. The motivation is that different IR systems will complement each other, because they usually emphasize different query features when determining relevance and retrieve different sets of documents. In clustering, documents are clustered either before or after retrieval [10] Combining multiple retrieval results is certainly a practical technique for improving the overall performance of information retrieval systems. But the proposed system deals with query expansion which uses document fusion because it effectively merge the results of different ranking functions that are applied to a single collection as compared to clustering technique used in different papers.

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 to perform the search on the web. The original query is expanded using the terms extracted in the Information Extraction module, and the resulting query is executed in the web search engine.

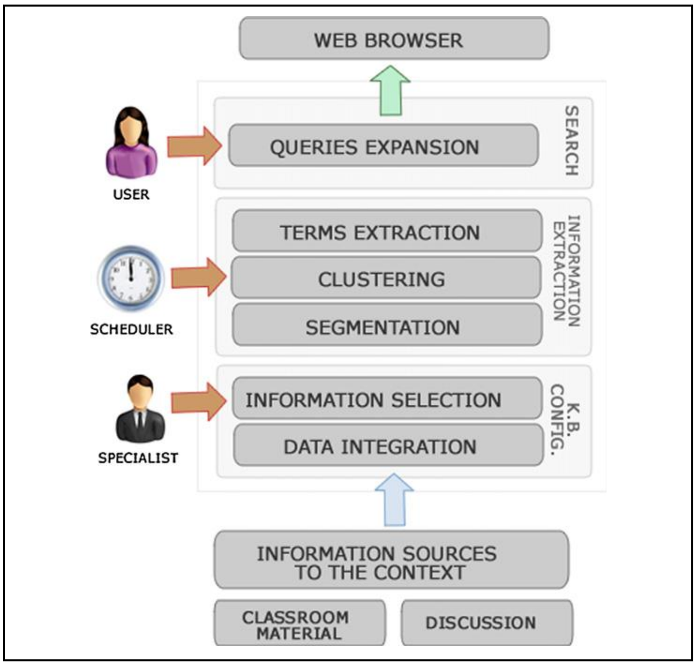

Fig. 1. Existing Architecture

In the proposed system, (shown in ***Fig. 2***) data is collected from the domain related and store in the database. The systems are collecting the domain based document e.g. *software engineering* based related document (This document is used for the indexing purpose). The user enter the query specific to the software engineering *for example: - what is analysis*. This query is going the system where the main operation is performing. The system uses this word as the terminology and searches this terminology in the database.

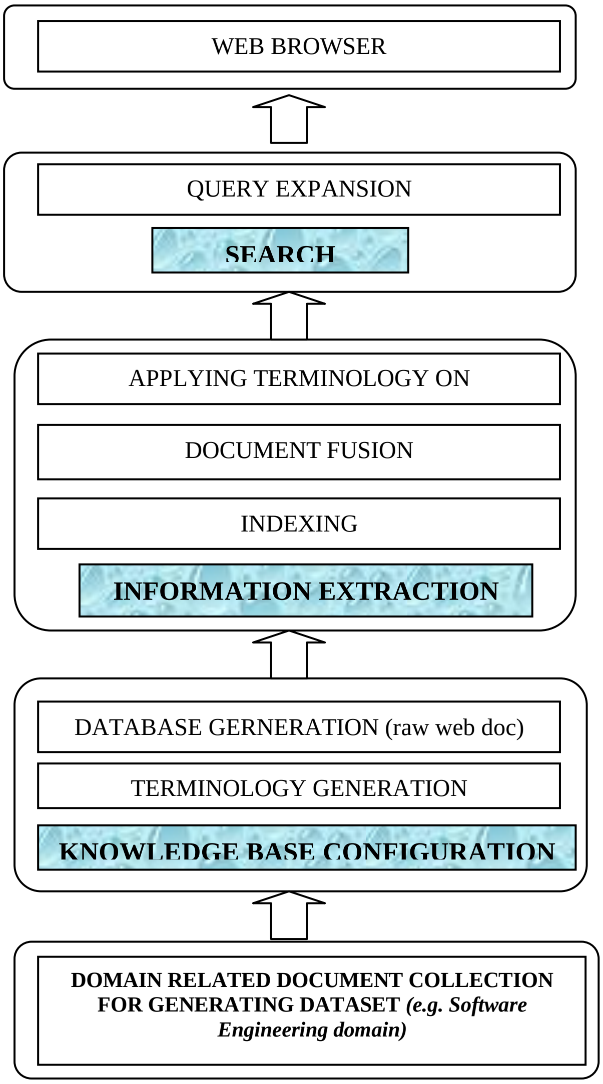
WEB BROWSER
QUERY EXPANSION
SEARCH
APPLYING TERMINOLOGY ON
DOCUMENT FUSION
INDEXING
INFORMATION EXTRACTION
DATABASE GERNERATION (raw web doc)
TERMINOLOGY GENERATION
KNOWLEDGE BASE CONFIGURATION
DOMAIN RELATED DOCUMENT COLLECTION FOR GENERATING DATASET (e.g. Software Engineering domain)

Fig. 2. Proposed Architecture

After searching the document using the data extracting technique the system extract the most relevant concept form that and created the document. Indexing and data-fusion techniques combine the rank lists of multiple document retrieval systems with the aim of improving search coverage and precision. Document fusion can effectively merge the results of different ranking functions that are applied to a single collection. After extracting the document the system will fused this concept using the data fusing technology. This fused document is then indexing using the BM25 model. BM25 is a retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document. The indexing is based on the score. The system then perform second main function i.e. expansion. The system expands the user query using the indexing page. It removes the stopped word and all unnecessary word and gives the first result. The user can select the second option for the query expands, the query is again is expanded which gives user better search result.

In the proposed system information retrieval is done by two main steps as ***(a) Conceptual Document Indexing*** and ***(b) Context Sensitive Information Retrieval.***

(a) ***Conceptual document indexing*** will tell How to extract key concepts from multiple terminologies? Suppose, given a collection of documents and n terminologies used for indexing, first extract concepts from each document D using a particular terminology Ti, i.e., it will obtain n lists of concepts for document D. Then it need to fuse n concept lists to obtain a final list of unique concepts representing various subjects matters of document D. Here purpose is to select the best concepts issued from several terminologies by means of voting scores assigned to candidate concepts. For this purpose, the propose system combine rankings of the extracted concepts from each document using their matching scores and/or their ranks. **(b)** ***Context sensitive document retrieval*****:** the document retrieval aims at matching the user's query to documents in order to retrieve a list of results that may satisfy the user information need.

In this system work, probabilistic BM25 term weighting model is used to rank documents, which are expanded with extracted concepts using multiple terminologies, with respect to a user query, where the relevance score of a document D for a query Q is:

$$Score(D,Q) = \sum_{t \in Q} w \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} + k_2 \,|\,Q\,|\, \frac{avdl-dl}{avdl+dl}$$

$$K = k_1((1-b)+b\frac{dl}{avdl-dl})$$

- k1, k2, k3, d: parameters
- qtf: query term frequency
- dl: document length
- avdl: average document length

The above equation is BM25 model equation which evalute the query and result is a list of documents, sorted by their similarity to the query.

E.g.	doc1	0.67
	doc2	0.65
	doc3	0.54
	…	

And highest query results match will tell the document required / demanded by the users.

Advantages of proposed system: It's a Domain Specific; also domain can be added in proposed system. In the proposed system most popular BM25 term ranking model is used. It provides two time query expansion. Instead of clustering and segmentation, document fusion and indexing is used in the proposed system to improve the retrieval effectiveness. Also Bose Einstein Statistic weighting model is used for checking search result this will tell how much search result is better.

V. 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. Due to which conceptual document indexing is carried out with the best result for the expanded queries. This work will give user best searching result by query expansion.

References

[1] "Contextual web searches in Facebook using learning materials and discussion messages" João Carlos Prates, Eduardo Fritzen, Sean W.M. Siqueira, Maria Helena L.B. Braz, Leila C.V. de Andrade, 2012 Elsevier Ltd. All rights reserved

[2] Fritzen, E., Siqueira, S. W. M., & Andrade, L. C. V. (2011). An agent-oriented system for contextualized web queries. In IADIS WWW/Internet 2011 (ICWI 2011), 2011, Rio de Janeiro. Proceedings (Vol. 10, pp. 479–483). Lisboa: IADIS Press.

[3] " Contextual Query based on Segmentation and Clustering of Selected Documents for Acquiring Web Documents for Supporting Knowledge Management" by João C. Prates UNIRIO, Sean W. M. Siqueira UNIRIO http://aisel.aisnet.org/amcis2011_submissions

[4] Sumathi, C. P., Valli, R. P., & Santhanam, T. (2010). Automatic recommendation of web pages in web usage mining. International Journal on Computer science and Engineering (IJCSE), 02(9), 3046–3052.

[5] S. Abdou, J. Savoy, Searching in Medline: query expansion and manual indexing evaluation, Information Processing Management 44 (2008) 781–789.

[6] The role of manually-assigned keywords in query expansion Kazem Taghva *, Julie Borsack, Thomas Nartker, Allen Condit Information Science Research Institute,University of Nevada, Las Vegas, NV 89154-4021, USA [7] Satish Gupta, "A White Paper on Home Networking", Wipro technologies, Bangalore.

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

[8] Bhogal, J., Macfarlane, A., Smith, P. (2007) A review of ontology based query expansion, *Information Processing andManagement*, 43 (4), July, 2007

[9] X. Zhou, X. Zhang, X. Hu, MaxMatcher: biological concept extraction using approximate dictionary lookup, in: Proceedings of the Pacific Rim International Conference on Artificial Intelligence, 2006b, pp. 1145–1149.

[10] Improving the Effectiveness of Information Retrieval with Clustering and Fusion Jian Zhang, Jianfeng Gao, Ming Zhou, Jiaxing Wang, Computational Linguistics and Chinese Language Processing Vol. 6, No. 1, February 2001, pp. 109-125 © Computational Linguistics Society of R.O.C.