IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 10, December 2014.
www.ijiset.com
ISSN 2348 – 7968

A Contextual Information Retrieval Using Query Expansion

Rushali Patil[1], Pramod Ganjewar[2]
[1] Department of Computer Engg, MIT AOE,Pune University,
Pune, Maharashtra 411015, India

[2] Department of Computer Engg, MIT AOE,Pune University,
Pune, Maharashtra 411015, India

Abstract
A contextual information system is developed by expanding original query with additional terms identified by global analysis to improve web search results by interactive query expansion. A context is any type of information which exhibits user's intention behind search. This contextual web search is applied to web based learning environment. In this a contextual web query tool is built to provide better search results than the usual traditional search approach. The contextual information is obtained from shared content (learning material provided by the teacher subject expertise) which is available in the learning platform and is used in a query expansion approach to get more suitable documents for the learning activities.
This work is a shift from query-centered IR to context-centered IR which aims to close a gap between user queries and resources available on WWW.
Keywords: *Context, Contextual Information Retrieval, Query Expansion, Web Search.*

1. Introduction

Information retrieval system (IRS) stores and accesses relevant information as per the user intention for search. IRS represents documents and queries as a set of weighted words, and similarity between them is found to produce ranked list of documents which are assumed to be relevant to the user query. Document and queries are represented using various models like the vector space and the probabilistic models. In this models the specific user's search intention and actual topic of the search is not taken into consideration. That is query based IRS ignores the user's search context and returns the similar results for the same query when entered by the different users [7]. In this type of IRS, user must have domain knowledge and interpret meaning of terms according to search goal.
Context is considered as any piece of information that can be used to characterize the situation of an entity. The contextual web is based on a user's intention behind search. The can be modeled from various information like the information on the page he is searching and the user's behavior against those pages. Once you understand the user's context, you can be more helpful. So, contextual technologies helps in improving web search results of query centric IR[13]. Context based IR (CIR) system is an improvement over traditional query based information retrieval (IR) system. In this system context or relevance of search goal is found before searching for desired documents (resources). This context of search goal is augmented by modifying original query with additional terms. These added terms are considered as query's context. The contextualized IR systems learn and predict in advance what information a searcher needs, learn how and when information should be displayed, present results by relating to previous information and to the tasks the user has been engaged in and decide who else should get the new information[13].

2. Related Work

The Connor and Limbu [3] study implemented and evaluated contextual retrieval system in which web search results improved using query enhancement and additional relevant information by taking users context into account through user contextual profile. The user profiles are modeled with the help of implicit (i.e. browsing and typing) and explicit (i.e. explicit rankings, inputs and instructions) data. For every user separete contextual profile is maintained to save contextual data and through knowledge base these profiles are shared among other users.

The approach [4] proposed tries to close the gap between users query and structured data. It mines query logs to find entity synonyms from structured documents which aim to gain user search intension. In this work[5], in a dynamic and automated fashion users historical queries are organized into groups which helps in functioning of different search engine components and also facilitate applications, such as query suggestions, ranking of results, query adaption, sessionization, and collaborative search. How to obtain, maintain and represent appropriate information about users multiple interests and how to use these acquired information to produce personalized search results are challenges to this approach.

This technique [1] involves users to modify query. Learning documents are utilized to get extra related information about user query. Text mining techniques has been applied on these shared documents to retrieve various subjects included in those documents. Distinct terms are presented to user for altering the query and thereby

ISSN 2348 – 7968

improving the results. In this approach [2], context has been modeled using a domain specific document which helps to find additional terms more relevant to users search goal. These retrieved terms are used for automatically expanding seed query.

3. Proposed System

The proposed system consists of three modules: Context Modeling Module, Term Selection Module, Search Module.

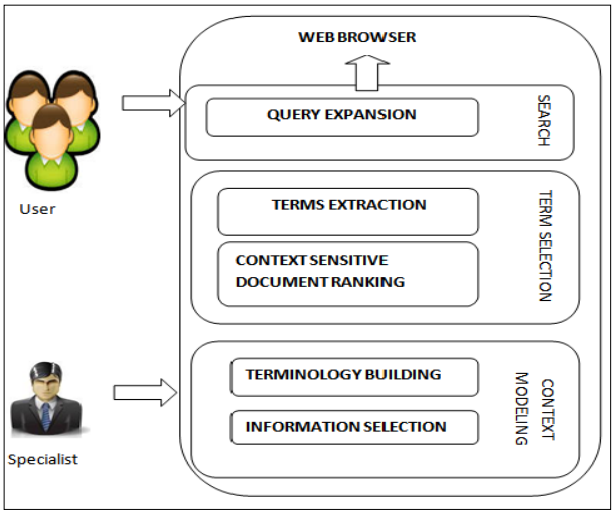

Fig. 1 Proposed System

3.1 Module I: Context Modeling Module

This module builds terminologies for specific domain and constructs knowledge base. In this module context of the future search is defined. This module consists of two sub modules namely: i) Terminology building and ii) Information Selection.

1) Terminology Building

Terminology is a set of all specialized terms that are representative of particular domain or subject. It helps in clear understanding of that domain. Hence a specialist in that domain builds terminology for a given subject and will further be used to classify selected resources. Terminologies are built for different subjects. In addition this terminology can be applied for automatic performance evaluation of IR. Based on terminology web search result can be categorized into relevant and irrelevant documents.

2) Information Selection

Existing educational resources such as files (articles, book chapters, publications in general) are used to model domain context. Terminologies from terminology building module used to categorize these resources before further processing of documents.

A domain specialist must select the contents that are good representatives of the domain from all available information sources. With the help of these content representatives (information resources) the system can work with multiple contexts. If the document shared as a source of context and includes terms from multiple terminologies then that document is included in all subjects.

3.2 Module II: Term Selection Module

The goal of the term selection module is to recognize the main terms of all the contextual information gained from the context modeling module, and to present a list of (additional) terms for the search module. This module comprise of further two sub modules: Document ranking and Term extraction.

Document ranking sub module ranks existing documents against query terms using BM25 algorithm for finding contextual information. Term extraction sub module analyses ranked documents and top 3 documents are selected as most relevant to query and then using these documents to find most relevant terms.

3.3 Module III: Search Module

This module search for original query and expanded query via web browser. It receives the keywords to perform the search on the web. The terms extracted in the Term Extraction module is used to expand original query and finally resultant query is executed in the web browser. All extracted terms (generated by the term extraction) are presented as a suggestion to user. As per interest the user has to select the terms to be augmented to the original query for achieving the query expansion.

4. Results

The system has been tested by users for different queries in software engineering and database system domain and has collected following results. While testing the system first 10 web search results (links) are considered for evaluation and search results have been analyzed using binary value. The value 0 indicates irrelevant document and 1 means relevant document. Precision and search length are two parameters calculated using two approaches: i) with original query ii) expanded query. The system shows improvement in precision parameter as results returned by web search engine are more relevant to user's context. It also decreases the search length value.

ISSN 2348 – 7968

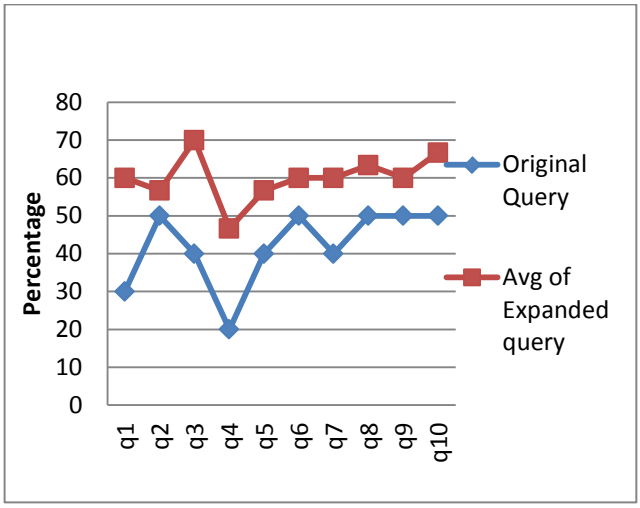

Fig. 2 Precision

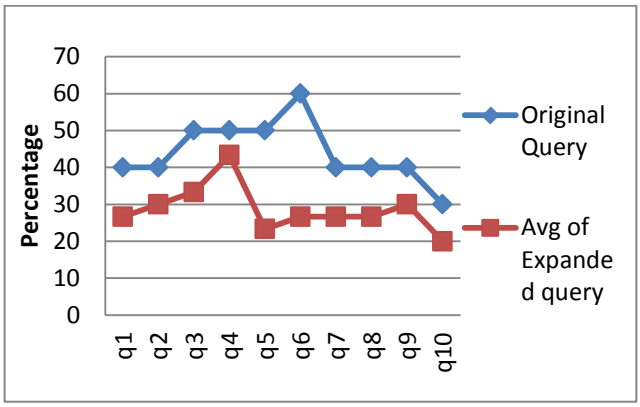

Fig. 3 Search length

5. Conclusions

The use of terminology, shared resources and query expansion can be considered to implement contextualization. Shared resources are compared with original query to rank document according to the higher similarity with query term. Then from top ranked documents terms are assigned weight to select most relevant terms for query expansion. The relevant terms are presented to user and then user will choose terms for formulating original query. These expanded query bridges the gap between users information need and traditional information search.

The use of global document analysis, BM25 ranking on existing resources in databases make search results more contextualized and therefore more useful to users. The results showed that the query expansion applied to a search engine in an educational environment can improve search results and the students were motivated to use the improved query functionality. The precision and search length parameter shows great improvement in result. Thus overall web search experience is improved.

References

[1] Joao Prate and Sean S.M. Siqueira,"Contextual query based on segmentation and clustering of selected documents for acquiring web documents for supporting knowledge management", American Conference on Information Systems,2011,1-9.

[2] Joao Prate and Sean S.M. Siqueira, "Contextual we searches in Facebook using learning materials and discussion messages", Computers in Human Behaviour,29(2),2013, 386-394.

[3] Andy Connor, Dilip Limbu, S. G. MacDonell and Russel Pears, "Improving web information retrieval using shared context", International Journal of Information Sciences and Computer Engineering, 2010,1(2),26-35.

[4] Tao Cheng, Hady W. Lauw, and Stelios Paparizos, "Entity Synonyms for Structured Web Search", IEEE Transactions On Knowledge And Data Engineering, 24(10), 2012, 1862-75.

[5] Heasoo Hwang, Hady W. Lauw, Lise Getoor, and Alexandros Ntoulas,"Organizing User Search Histories",IEEE Transaction on Knowledge and Data Engineering,24(5),2012,912-925.

[6] Bodo Billerbeck and Justin Zobe,l, "Document expansion vs Query expansion for Ad-hoc retrieval", Australasian Document Computing Symposium, Sydney, Australia,10,2005.

[7] Ian Ruthven, "Information retrieval in context", Information Retrieval Series,33,2011,187-207.

[8] Prates and Siqueira, "Using educational resources to improve the efficiency of Web searches for additional learning material", IEEE International Conference on Advanced Learning Technologies, 11, 2011, 563-67.

[9] Duy Dinh and Lynda Tamine,"Towards a context sensitive approach to searching information based on domain specific knowledge sources", Web semantics:Science,Service and Agents on the World Wide Web,12, 2012,41-52.

[10]Limbu, Connor,Pears and Stephen,"Contextual relevance feedback in web information retrieval", International Conference on Information Interaction in Context, New York USA,ACM,2006,138-143.

[11]Carpineto and Romano, "A Survey of Automatic Query Expansion in Information Retrieval", ACM Computing Surveys, 44(1)

[12]Fazli Can, Rabia Nuray and Ayisigi B. Sevdik, "Automatic performance evaluation of Web search engine", Information Processing and Management, 40, 2004, 495–514.

[13] Rushali Patil and Pramod Ganjewar, "Improving Contextual Web Searches through Query Expansion Using Domain Specific Resources", International Journal of Engineering Research & Technology (IJERT), IJERT, Vol.3(3), 2014, 2285-88.

ISSN 2348 – 7968

Rushali Patil Member of Army Institute of Technology, Pune since 2007 and Student(S) at MIT Academy of Enginnering from 2011-12. I am situated in Pune and born on 23rd of September, 1982. I have secured first class B. E. degree from Veermata Jijabai Technological Institute(VJTI) , Mumbai , Maharashtra in 2005 . Presently working as an Assistant Professor at Army Institute of Technology in Pune.

Pramod Ganjewar PhD Research Scholar, Sathyabama University, Chennai - 600 119, Tamilnadu, India Assistant Professor, Department of Computer Engineering, MIT Academy of Engineering, Dehu Phata, Alandi(D.), Tq. Khed, Dist. - Pune - 412 105, Maharashtra, India.