International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 3, March - 2014

Improving Contextual Web Searches through Query Expansion Using Domain Specific Resources

Rushali Patil[1]
1.Student, Department of Computer Engg.,
2.MIT AOE, Pune, India.

Pramod Ganjewar[2]
1. Assistant Professor, Department of Computer Engg.,
2.MIT AOE, Pune, India

***Abstract*- Information retrieval (IR) is a process which aims at storing and accessing relevant information as per the user needs. The main objective of an information retrieval system (IRS) is to present to the user the relevant documents in response to his queries,but this kind of search is usually performed using keywords and results into context less data. There is a mismatch between user intend for information and data presented to user for entered query. This work is one step towards better understanding of user context from available domain specific resources. The contextualized strategy for information retrieval (IR) can be built around user profile, query expansion and relevance feedback.**

This paper proposed a method that exploits new concept called terminology which is good representative of particular domain (or subject) and used to classify resources as relevant or irrelevant (non- relevant). This work focuses on improving results of context sensitive web search engine based on shared resources and query expansion technique. Most relevant terms are inferred from shared documents to formulate original query. These resources are ranked based on query submitted by user which brings most relevant document at the top in hierarchy. From top ranked documents, terms has been weighted to identify most related terms for query expansion. In addition the results of the query engine with and without the contextual information will be evaluated automatically without any interface of user.

***Keywords*- context, contextual information retrieval, query expansion, web search.**

I. INTRODUCTION

World Wide Web is a large repository of information available via internet. Size of WWW has been continuously increasing and this information is disseminated to a wider audience, due to which user has excess amount of data. This extraneous data causes information overload problem and gaining information resources relevant to an information need from a collection of information resources becomes difficult. The activity of retrieving relevant information resources as per users information need from available resources is called as information retrieval. After receiving query from user traditional query-centric IR executes it in search engines and results are presented to user. Traditional IR is like single solution to different types of problem i.e. it doesn't take into account user's information seeking behavior. User's information need is something more than entered terms (words). This something means relevancy in search and it is found using context of the terms. This type of IR ignores several implicit factors about the user and the search context (e.g. time, location, task, expertise, interaction) are ignored and can be considered for optimization of IR performance. If user fires a query "java" to download a software, then IR returns results including "java coffee shop", "java island", "java software", etc. Thus context of search is unnoticed which is important to produce relevant documents. Thus relevant documents will appear in first one or two pages and remaining pages aren't of any use. Information seeker must remember of the key terms to identify resources [1]. This technique has following limitations:

1. User needs to have domain information.
2. Terms have totally different meanings and interpreting meaning of term as per search goal is user's responsibility.

Alternative to traditional query centric information retrieval (IR) is context sensitive IR (CIR). Context sensitive means relevance of search goal is found before searching for relevant documents (resources). This relevance of search goal is added by expanding original query with few terms. Now these augmented 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. Context is any information which can be used to identify the situation of any entity. An entity is a person, a place or an object that is considered relevant to the interaction between a user and an application, including the user and applications themselves [2]. The CIR is based on three major themes: user profile modeling, query expansion, and relevance feedback. In this work predefined resources are used to decide context of user query and additional terms recognized as relevant to original query are used to expand query. The goal of query expansion techniques is to improve the way search engines cope with user queries.

This paper is organized as follows. In Section II an overview of related works is given. We then explain the importance of context and contextual information system in Section III. Section IV introduces proposed system and its module in

detail. Automatic relevance calculation method for results returned by proposed systems is presented in Section V. Conclusions are discussed in Section VI.

II. RELATED WORK

The Connor and Limbu [3] study implemented and evaluated contextual retrieval system. In this work implicit (i.e. browsing and typing) and explicit (i.e. explicit rankings, inputs and instructions) data has been utilized to provide relevant information to user. An individual contextual profile is maintained to store this data and then these profiles are shared among other users through knowledge base. The system has integrated two levels of recommendation support: 1) suggestion of similar terms and concepts for refining a query 2) recommending relevant pages previous visited by user when a shared contextual knowledge base is enabled.

The study reported in Cheng and Lauw [4] has focused on improving structured web search. Nowadays resources available on web are structured data (e.g., movie showtime of a specific location). Hence there is often a mismatch between the data(i.e how it is created and presented to user) and the web queries (how different users try to retrieve them). In this work entity synonyms are found over structured data by mining query logs. Due to this approach number of web pages returned (or covered by search engine) for a user query is increased.

Hwang and Lauw [5] discussed about query reformulation and click graph. To fulfill user information need on the web, search engines keep track of their queries and clicks while searching online which is organized to build relevant information for that user.

Bodo [6] investigated the use of document expansion as an alternative, in which documents are augmented with related terms extracted from the corpus during indexing, and the overheads at query time are small. This work has proposed and explored a range of corpus-based document expansion techniques and compare them to corpus-based query expansion on TREC data.

Prates and Siquiera [1] considered internet as a rich source of information and can be used in educational environment. This work uses lecture notes and other resources to model context of students studying same subject. On these shared resources information extraction techniques have applied and relevant terms are identified to expand user entered query.

Prate and Siqueira[2] enhanced above[1] work by utilizing shared documents and discussion messages posted in a social network which is used for collaborative learning.

III. CONTEXT INTEGRATION INTO IR

The Information Retrieval (IR) process performance is depending on the context in which a search takes place. The searcher's interaction with the IR system, his expectations and his decisions about the documents he retrieves are influenced by context. Hence, identifying what features are important in a searcher's can be helped to design more useful and successful IR systems. In general, the user profile, the activity developed at the time, search history or information obtained through sensors represented in particular format can be used to model some kind of context, which can be the domain of some knowledge.

The normal process of documents retrieval can be modified using context model, providing new terms for expanding the original query or serving as a basis for reclassification of the order of relevance of initial set of retrieved documents. Context model can be built automatically or manually. Different uses of context are listed below:

1. Predicts what information end user need.
2. Indicates interrelation between information.
3. Decide recipient of similar type of information.
4. Learns what information user need by observing his reactions to presented information.

The Resources available on World Wide Web are huge and subject to different search contexts that need to be considered while searching for information need. The key issue in IR is: how to capture and how to incorporate contextual information in the retrieval process in order to increase the search results. In [10], contextual retrieval is stated as building contextual knowledge bases by capturing user's implicit and explicit data (i.e. combining user context into a single framework) in order to give the most suitable answer for user's query.

Thus, contextual IR focuses on two steps to bridge gap between available resources and users query:

1. Defining the search context i.e. context of user information needs.
2. Modifying the search by taking context into account in the information selection process.

IV. PROPOSED SYSTEM

A. Context Modeling

In this module context of the future search is defined [2]. 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 [9]. It helps in clear understanding of that domain. Hence building terminology for a given subject is accomplished by specialist in that domain. These terminologies 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 [1]. Any source of information containing textual content and information, which represent the relevant topics in the domain context, can be used for this purpose. These resources will be categorized using terminologies from terminology building module 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.

B. Term Selection

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 two components: i) Document ranking, ii) Term extraction

This module introduces a method to evaluate the (selected or predefined) document content for finding contextual information. From ranked documents, top 10 documents are selected as most relevant to query and then using term

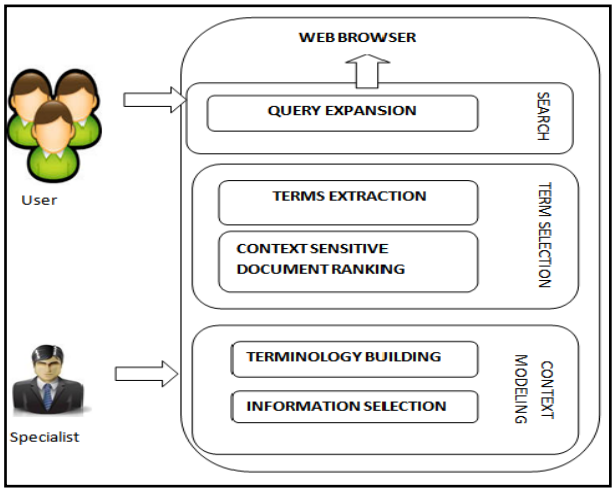

Fig. 1: Proposed System Architecture

selection value (TSV) candidates terms are ranked. Terms having higher TSV value are selected for query expansion. This module applies following algorithm:

Algorithm:
Input: Q a query from user
Output: T a set of context sensitive terms
1: rank documents d against q
2: select top 10 documents R as local set
3: rank candidate terms using *TSV*
4: add top |E| = 25 terms to T

1) Context Sensitive Document Ranking

Document similarity with query terms is measured by applying BM25 ranking function. It is used to rank matching documents from context modeling module according to their relevance to a given search query.

2) Term Extraction

We use the *term selection value* in our experiments for ranking terms [6].

$$TSV_t = \left(\frac{f_t}{N}\right)^{f_{r,t}} \binom{|R|}{f_{r,t}}$$

where,

f_t is the number of documents in the collection in which term t occurs in,

N is the total number of documents in the collection, and

$f_{r,t}$ is the number of the |R| top ranked documents in which term t occurs.

3) Search Module

The search module 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.

V. AUTOMATIC RELEVANCE CALCULATION OF WEB SEARCH ENGINE'S RESULTS

Relevance feedback method which involves human interaction for measuring the information retrieval effectiveness of World Wide Web search engines is costly and time consuming. In this work proposed system's performance is measured automatically. Original query and expanded query are submitted to web search engine. After this the top 10 search results returned by search engines. The content of these retrieved pages, if the pages are available, are downloaded and saved to build a separate document collection, or Web abstraction. In this process unreached-links are considered as useless resource. In this work terminologies are considered as appropriate representative of domain, hence can be used to judge the relevancy of document. The downloaded pages are all translated to plain text (ASCII) representation and non-word tokens, such as page structure information (e.g., HTML, pdf, ps tags), are deleted. After these preprocessing, documents are classified as relevant or non-relevant based on the terminologies built in context modeling module. Similarity matching between terms from terminologies and documents is performed by employing support vector machine (SVM). According to this similarity score documents will be ranked.

Three metrics are used to analyze the search results: precision, search length and rank correlation.

Precision is a metric commonly used in information retrieval and represents the fraction of retrieved documents that all relevant to user's information need. The different amounts of relevant information found in each document is not considered by the binary judgment of relevance(i.e. 0 for irrelevant and 1 for relevant document). The full precision metric considers the total amount of relevant information identified in the first 10 results, through the use of a Similarity score. The lower score implies that the result has no relevance and value the higher value indicates high relevance.

The search length is the second metric, which reflects the number of non-relevant documents that the automated system must evaluate until identifying a certain number of consecutive documents that are considered relevant. Therefore, lower values imply less effort to find relevant results. The search length is defined as the number of documents evaluated until two consecutive results were found with the value of relevance greater than or equal to three.

The last metric is rank correlation which intends to correlate the rank order assigned by the search engine and the automatic relevance judgment, where the results are sorted in descending order of relevance. The correlation between the relevance of search results and the ideal prioritization is higher means the search tool is more effective.

VI. CONCLUSION

This work has introduced the use of terminology, shared resources and query expansion to implement contextualization which in turn improves search results. Shared resources are matched 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 user's information need and traditional information search.

This works also introduced automatic performance evaluation of context sensitive information system. For these terminologies from context modeling module is used to categorize relevant and non-relevant documents. Therefore user is no more involved in relevance judgments.

REFERNCES

[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.