Application for Web Data Extraction and Analysis

Preeti Rawat¹, Rajesh², Sarah Sayyad³, Suveni Surinder⁴ & Prof. S N Shelke⁵

¹,²,³,⁴ B.E. Computer Engineering Students, Sinhgad Academy of Engineering, Pune.
⁵Professor, Sinhgad Academy of Engineering, Pune

Abstract: In web search applications, the World Wide Web (WWW) is a huge resource for people. This resource uses search engines to search the information. The web uses search engines to search different kinds of information where queries are submitted to search engine. Queries submitted represent needs of the users. Sometimes they may not exactly be representing the actual objective of user. As there are many ambiguous queries and different users may want to get information on different aspects when they submit the same query. User search objective is information on different aspects of a query that user groups want to obtain. Information need can be defined as a particular desire to obtain information to satisfy ones need. User’s search goals can be considered as the clusters of information needs for a query. Analysis and inference of user’s search goals have many advantages in improving user experience and search engine relevance.

1 INTRODUCTION

In this project, we aim to discover the number of distinct user search goals for a query and mapping each goal with some keywords automatically. We first propose a novel approach to figure out user search goals for a query by clustering our recommended feedback sessions. Then, we propose an innovative optimization method to map feedback sessions to pseudo-documents which can satisfactorily reflect user information requirements. At last, we cluster these pseudo documents to infer user search goals and relate them with some keywords.

The different users have different search objectives when they submits query to the search engine. As there are so many ambiguous queries and different users may want to get information on different aspects when they submit the same query. In fig 1 we can see that user has entered “the dawn” as a query which is a homonyms word. One can want information of Spacecraft and another may want to locate home page of Pakistan English newspaper. User search objective is the information on different aspects of a query that user groups want to obtain. There are mainly two aspects for improving search engine relevance 1) inference of user search objective and 2) analysis of user search objective. Both aspects have some advantages summarized as follows.

First, we can rearrange web search results acquired according to user search objectives. Rearranging or restructuring can be done by grouping the search results with the same search objective. By using this, we can easily find out users with different search objectives. And also we can find what exactly user want. Second, we are representing user search objectives by some keywords. These keywords can be utilized in query recommendation. By using recommendation, we find out the suggestions about queries which can help users to form their queries more detail. Third, the search objectives are distributed for the use of reranking of documents.

Figure 1.1 Example of different user objectives for “the dawn” query
[Ref- www.google.com]
2 IMPLEMENTATION
The main motivation of our system is to identify search objective and return search results within minimum time. In feedback session it keeps more number of URLs as compared with previous method. New framework makes use of Feedback session and Hierarchical clustering.

2.1 Databases

2.1.1 WordNet
WordNet is a lexical database for the English language. It clusters English words into sets of synonyms called synsets, which provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. WordNet can thus be viewed as a combination of dictionary and thesaurus. While it is accessible to human users via a web browser, its primary use is in automatic text analysis and artificial intelligence applications. The database and software tools have been released under a BSD style license and are available free of cost for download from the WordNet website. Both the lexicographic data (lexicographer files) and the compiler (called grind) for producing the distributed database are available. WordNet includes the lexical categories nouns, verbs, adjectives and adverbs but ignores prepositions, determiners and other function words.

2.1.2 Database
In our system, database will store all the information related to registration, authentication and modification. It is an important part of the architecture as the first process that is login starts with the database.

2.1.3 SQL Server
SQL Server is an open-source application server project started by Sun Microsystems for the Java EE platform. SQL server is now sponsored by Oracle Corporation. SQL Server is the reference implementation of Java EE and as such supports Enterprise JavaBeans, JPA, Java Server Faces, JMS, RMI, Java Server Pages, servlets, etc. This allows one to create enterprise applications are portable and scalable in nature, and that integrate with legacy technologies. Optional components can also be installed for additional services.

2.1.4 Web Services
All the web services related to the SQL Server will be included in our implementation of the project. Services are like developing of web page.

2.2 MODULE AND ALGORITHM
The main motivation of our system is to identify search objective and return search results within minimum time. In feedback session it keeps more number of URLs as compared with previous method. New framework makes use of Feedback session and Hierarchical clustering.

2.2.1 A Proposed Framework
The figure 2.1 shows the framework of our approach. The proposed system framework is an enhancement to techniques introduced in . The main motive of proposed system is to identify search objective and return search results within few time. In feedback session it keeps more number of URLs as compared with previous method. New framework makes use of Feedback session and Hierarchical clustering.

Figure 2.1: Proposed System Model (Framework)

Proposed framework consists of different parts as follows –

Part 1: In the first step queries are submitted to the search engine and all these queries are stored into Click through logs. Different users have different aspects for searching information. Feedback sessions of a query are first extracted from user click-through logs. In the feedback session different types of information is stored like Click sequence, Date/Time, IP address. The feedback session is useful for inferring user search objective. In the proposed system we can consider number of feedback session for analysis. These feedback sessions are mapped to the pseudo-documents. Pseudo-documents consist of keywords which represent the user’s information need. By using these pseudo-documents we can easily find out the user search objectives inferred by clustering. Initially we do not know the exact number of user search objective so that several different values are tried. From these values optimal value will be determined by the feedback.
Part 2: In this the clustering of the feedback session takes place. One can cluster the retrieved results but in the proposed system feedback sessions are clustered. Clustering of the feedback session is more difficult than the clustering of the retrieved results. Finally, we evaluate the performance of retrieved results by our proposed evaluation criterion CAP. And the evaluation result will be used as the feedback to select the optimal number of user search goals in the part 1. In this paper focus is on the feedback sessions, pseudo documents, hierarchical clustering. For identification of user search objective following procedure we have to follow:

I. Feedback session:
When user submits the query to the search engine we do not know about the exact need of user. So for the identification of user search objective we are maintaining feedback session. Feedback session is the series of queries and some clicked search results. In the previous method it keeps the record of limited URLs. But in the proposed method number of URLs is increased. By using feedback session we can easily identify user search objective. In this we are considering feedback session for only single query. Therefore for the single query feedback session is also single. The proposed framework consists of feedback session with both clicked and unclicked URLs and they ends with the last URL which was clicked in single session. Before the last click all previous URLs have been scanned and they are evaluated by users. Feedback session consists of 0 which indicates that corresponding URL is unclicked. In feedback session only three URLs are shown. In that one URL is clicked and two URLs are unclicked. Clicked URL indicates what exactly user needs. Unclicked URL shows that user doesn’t want the information related to this URL. Therefore, for identification of user search objective it is best way to analyze the feedback session then the analyzed search results or clicked URLs.

II. Mapping of feedback session to pseudo document:
There different methods for representation of feedback session. One of the methods is known as Binary Vector Method to represent feedback session. When the query the sun submitted to the search engine 0 represents unclicked in the click sequence. For example, binary vector [0110001] can be used to represent feedback session. In that 1 represents clicked and 0 represents unclicked. The binary vector method has disadvantage that doesn’t give enough information to identify user search objective.
New method is proposed for representation of feedback session. In this method the feedback session is mapped to pseudo documents. Pseudo documents consist of keyword to determine whether document can satisfy their need. Hence pseudo documents can be used to identify user search Objective.

III. Algorithmic strategy for proposed framework:
a) Generate pseudo documents of URL from feedback session.
b) Pseudo document contains all the keywords from given web pages title and description.
c) Also all the stop words and stemming words are removed.
d) Apply K-means clustering algorithm to form a group of relevant keywords from pseudo documents so that each cluster represents one user search objective.
e) Next step is to organize words from one cluster into topic-subtopic hierarchy by using clustering.
that clustering feedback sessions is more efficient than clustering search results or clicked URLs directly. Moreover, the distributions of different user search goals can be obtained conveniently after feedback sessions are clustered.

d) We propose a novel optimization method to combine the enriched URLs in a feedback session to form a pseudo-document, which can effectively reflect the information need of a user. Thus, we can tell what the user search goals are in detail.

4 CONCLUSION

In our method we have resolved the problems associated with feedback session record. Feedback session can record limited number of URLs. So that user can analyze few URLs. In this case we have increased the size of feedback session. So that user can analyze more number of URLs. In pseudo documents keywords are present which are clustered according to clustering. We used hierarchical clustering for searching topic-subtopic wise. From this method user can easily find out his/her information need within small time. We studied and implemented feedback session and mapping of these feedback session to the pseudo documents. Finally we also implemented performance method to evaluate search results. For measurement of the performance classified average precision (CAP) method is used .This approach is used to improve searching within minimum time. The system framework is useful and feasible to be used with real world search systems. It will help users to search information more precisely.

5 ACKNOWLEDGEMENT

This research paper is made possible through the help and support from everyone, including: parents, teachers, family & friends. We would like to show our gratitude towards Prof.S N Shelke for his utmost support and encouragement. He kindly read our paper and offered invaluable detailed advices on grammar, organization, and the theme of the paper. We would also like to thank Prof.B B Gite to reproof the paper as well as all the other professors who have taught us about the techniques we’ve used in our project. The product of this research paper would not be possible without all of them.

3 BENEFITS

a) Users having different search goals can easily find what they are searching.
b) Recommended queries can help user to form their query more accurately.
c) Used in application like re-ranking web search results. We propose a framework to infer different user search goals for a query by clustering feedback sessions. We demonstrate
REFERENCES

[1] A Novel Optimization Algorithm based on Reinforcement Learning, Janusz A. Starzyk, Yinyin Liu, Sebastian Batog


