

NOVEL APPROACH FOR INFORMATION RETRIEVAL

Sneha A. Taksande¹, Prof. A. V. Deorankar²

¹PG Scholar, Department of Computer Science and Engineering,
Government College of Engineering, Amravati, Maharashtra (India)

²Associate Professor, Department of Information Technology,
Government College of Engineering Amravati, Maharashtra (India)

ABSTRACT

Nowadays, all information available on World Wide Web present in digital form. With the gradual increase in the amount of information in the World Wide Web, there is a need for a much efficient techniques for Web Search. Sometimes the traditional keyword matching as well as the standard statistical techniques are insufficient to retrieve more relevant Web Pages. Users expect required information to be retrieved in return of simple short query with generic web search on huge heaps of information but there is great difficulty in retrieving relevant information according to user preferences. So, we need to enhance the power of web search to retrieve relevant information. Different Information retrieval (IR) techniques are available there. So this paper is an attempt to provide different methods to retrieve the most relevant information from such a huge collection that satisfied the users need. Thus it will represent the collaboration of various methods for information retrieval.

Keyword: World Wide Web, Information retrieval, Web Search, Information retrieval (IR) techniques.

I. INTRODUCTION

In today's time millions of people around the world uses Search engines. The term *search engine* denotes not only well-known commercial Web search engines such as Google, Yahoo, and Bing but also a wide range of search systems that are part of major Web-based applications such as email and social networks. With the inception of World Wide Web, the volume of data present on the internet is tremendous which makes it quite difficult for the user to navigate through this enormous amount of data. The need for the development of an automated system that can extract the required information becomes urgent as users struggle to navigate through this wealth of information. Information mismatch and irrelevant document result can be the two major fundamental problems with traditional web search. Suppose if one user is a computer engineer then he will thinks 'mouse' as computer peripheral device on the other hand if another user is a biological researcher then he will thinks 'mouse' as mammal. So the Semantic meaning of term 'mouse' can be different according to individual user. This leads to the query ambiguity. Thus a lot of research is focused on improving the ease of retrieval of the data. This paper presents the various information retrieval methods that consider different aspects of users to deliver relevant search results. This includes the methods based on personalization that takes into

account the users profile, search history and his behavior on web search. The keyword based search engines that rely on keyword matching usually return too many low quality results and also does not help in solving the ambiguous queries issue encountered. So to overcome this challenges clustering method can be used to overcome the ambiguity of the user query. Further personalized clustering technique can be used to manage the size of the cluster.

A search engine is graded by two main parameters namely relevance of information retrieved and the response time. An optimal search engine is said to have a very high efficiency if it furnishes information that is relevant to the query and has a very high response time. A Differential Adaptive PMI retrieval method is also formulated with varied thresholds for recommending the Web Pages based on the input query. This methodology yields the better accuracy.

II. RELATED WORK

For evaluating the performance of any website, information of web users plays a very important role. Several tools are available commercially in order to access this information. In paper [5] Query logs are considered as one of the most valuable tool for search engine optimization. Every search engine maintains a log of what users search on to it including user ID, query, clicked URLs, rank of URLs and time of access. Thus, this huge amount of information from query logs can provide an insight into user browsing behavior and users' information needs.

A new approach to automatic indexing and retrieval is describe in [9] which is designed to overcome a fundamental problem that plagues existing retrieval techniques that try to match words of queries with words of documents. The problem is that users want to retrieve on the basis of conceptual content, and individual words provide unreliable evidence about the conceptual topic or meaning of a document. So to overcome this problem, automatic indexing and retrieval approach is designed.

An innovative framework is proposed for yielding the most relevant web pages in[4]. A novel strategy called as Differential Adaptive Pointwise Mutual Information is proposed for computing the semantic heterogeneity which is one of the primary contributions to work of this paper. The query words are used for extraction of the relevant URLs from the URL repository. From the URL structure, the keywords and content words are extracted. The semantic similarity is computed between the keywords and the content words to obtain a feasible word set.

This paper [6] gives the personalized search system which uses the alternate query generator to capture all the senses of the main query and assists the user with the alternate queries. Further personalization based upon users profile, click history and last action performed by the user is used to improve search results.

III. METHODS FOR INFORMATION RETRIEVAL

3.1 Relation-Based Information Retrieval

It is worth observing that statistical algorithms are applied to “tune” the result and, more importantly, approaches based on the concept of relevance feedback are used in order to maximize the satisfaction of user's needs. Nevertheless, in some cases, this cannot be sufficient. In a traditional search engine like Google, a query

is specified by giving a set of keywords, possibly linked through logic operators and enhanced with additional constraints.

Let us assume now that the user specifies the keyword “Rome,” and he or she then selects from the pull-down menu one of the possible concepts such as Destination or City. A second keyword “hotel” is then added, choosing accommodation as the associated concept. In general, there is no way to state which was the relation in the user’s mind between those two concepts.

Now, let us consider a set of interpreted pages containing keywords “Rome” and “hotel” and associated concepts Destination and Accommodation. A traditional search engine like Google would return both pages without considering the information provided by the semantic mark. On the other hand, a semantic search system would take into account keyword concept associations and would return a page only if both keywords (synonyms, homonyms, etc.) are present within the page and they are related to associated concepts. Finally, a relation-based search engine like the one presented in would go beyond pure “keyword isolated” search and would include these pages in the result set only if there exist enough relations linking considered concepts.

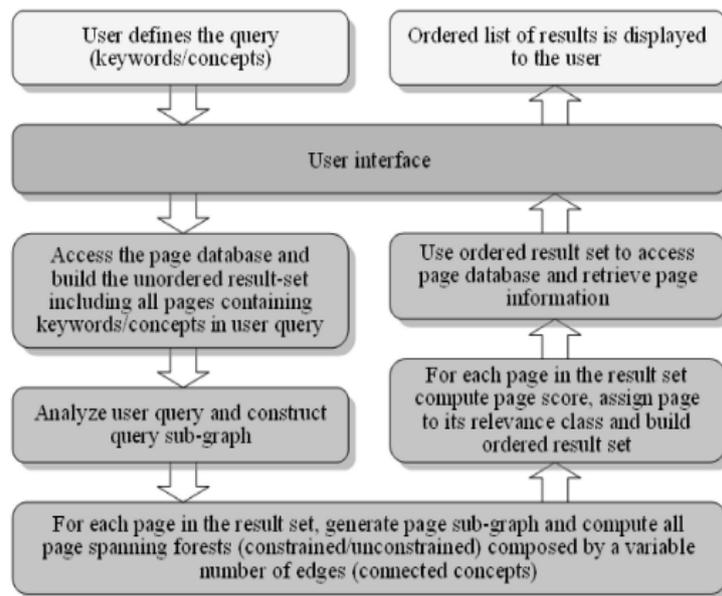


Fig. 1. Workflow from query definition to the presentation of results.

3.2 Differential Adaptive Pointwise Mutual Information Retrieval

An innovative framework for yielding the most relevant web pages is proposed. For computing the semantic heterogeneity a novel strategy called as Differential Adaptive Pointwise Mutual Information is proposed. The query words are used for extraction of the relevant URLs from the URL repository. From the URL structure, the keywords and content words are extracted. The semantic similarity is computed between the keywords and the content words to obtain a feasible word set. Additionally, the semantic heterogeneity is computed between the query words and feasible word set differentially to re-rank the URLs before they are yielded to the user. A higher precision, recall and accuracy are achieved for this methodology.

The architecture of this system is depicted in Figure 2. The query that is input from the user is preprocessed at first. Query Preprocessing involves parsing and tokenization of the multiword query. Furthermore, the query processing is subjected to stanching where the removal of stop words takes place and redundant words in the

query are eliminated. A query keyword set is formulated which comprises of unique query words. The proposed system architecture incorporates a URL Base which is a very large repository that houses a large volume of URLs. The URLs in the URL Base are collected from several web sources. There are individual URLs and also the Web Log Information of various users' which are aggregated together to constitute a URL base.

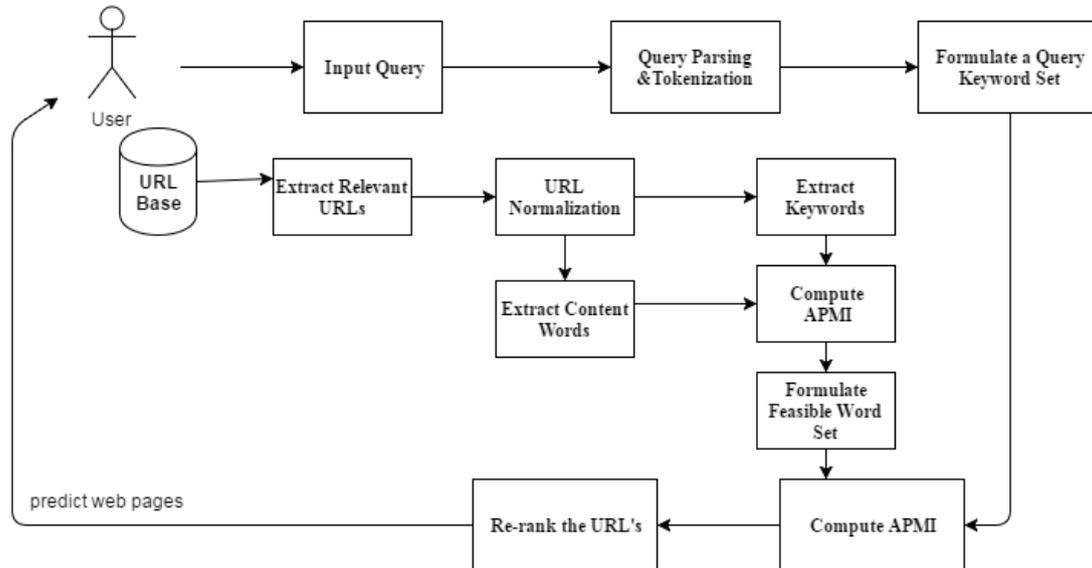


Fig. 2. System architecture for Differential Adaptive Pointwise Mutual Information Retrieval

3.3 Personalized Based Image Retrieval

The fast growing world of information technology and Internet made search engines serve as the main information portal for ordinary users. Sometimes search engines may return irrelevant search results so the users might often experience lack of efficiency and effectiveness. In such situations Personalized Web Search (PWS) gain importance. PWS is a general category of web searching technique that provides efficient search results and meeting individual user needs. Personalizing search can be based on user search histories, user interest and user's profile. Personalization is the process of providing right information to the right person at the right time. For personalization user interests needed to be studied which requires collection, analysis and accumulation of user data both general and personal.

The limitations of the existing personalized search methods can be lack of runtime profiling, lack of user privacy requirements customization and iterative user interaction for personalization is required. For minimizing the above limitations, user profile was built hierarchically with user interests. User specified sensitivity for any topic is hidden in the generalized profile. Personalization can be done in two phases: offline and online.

In the offline phase, a user registered with the system was provided entry and the user profile was built hierarchically by collecting information like name, ID, age, profession, etc. from the user. In this phase, the user could customize his/her interests and also the sensitive topics the user needs to hide. The user could rate his/her degree of sensitivity for the sensitive topics. Next the online phase will be performed. In the online phase, the user query was submitted. Then the query was mapped to the topic. The relevance of the topic was also considered. The last step of online phase was cost-based generalization, where a seed profile was generalized in cost based manner. Several algorithms were used to measure utility and risk

IV. CONCLUSION

Several methods for information retrieval are introduced in this paper. All these methods worked at their best. Lots of research is focused on enhancing the power of search engines to deliver the appropriate results according to users query. So this papers presents the collaboration of all these techniques that better works for retrieving the relevant information. It gives the technique for semantic search that is capable of exploiting concepts and relation between these concepts. Differential Adaptive Pointwise Mutual Information Retrieval that computes the semantic similarity between the querywords, keywords and content words differentially withheterogeneous thresholds is used to enhance the power of search system. It also includes the method that analyses user's behavior based on search intentions for information retrieval using the context search. Personalization is another tool that can be used to improve accuracy of search results based on user's search histories, user's interest and user's profile.

REFERENCES

- [1] JigarJadav, Andrew Burke, Pratik Dhiman, Michael Kollmer and Charles Tappert, "Classification of Student Web Queries," IEEE, 2017.
- [2] AvaniChandurkar and Ajay Bansal, "Information Retrieval from a Structured Knowledge Base," IEEE, 2017.
- [3] Sanjib Kumar Sahu, D. P. Mahapatra and R. C. Balabantaray, "Analytical Study on Intelligent Information Retrieval System Using Semantic Network," IEEE, 2016.
- [4] Gerard Deepak, J SheebaPriyadarshini and M S HareeshBabu, "A Differential Semantic Algorithm for Query Relevant Web Page Recommendation," IEEE International Conference on Advances in Computer Applications (ICACA), 2016.
- [5] ShipraKataria and PoojaSapra, "A Novel Approach for Rank Optimization using Search Engine Transaction Logs", International Conference on Computing for Sustainable Global Development, IEEE, 2016.
- [6] ShilpaSethi and Ashutosh Dixit, "Design of PersonalisedSearch System Based On User Interest and Query Structuring", 2nd International Conference on Computing for Sustainable Global Development, IEEE, 2015.
- [7] Shogo Kori, Yanjun Zhu, Koichi Yamaguchi, Satoru Takiguchi, YasufumiTakama, "Analysis of User's Behaviour Based on Search Intentions for Information Retrieval Using Search Engines," IEEE 2015.
- [8] Prakasha S, Shashidhar HR and Dr. G T Raju, "Structured Intelligent Search Engine for Effective Information Retrieval using Query Clustering Technique and Semantic Web," IEEE, 2014.
- [9] Ehsan Nowroozi, "Introduction to New Methodologies and Applications in Information Retrieval Indexing," 2nd International Conference on Mechanical and Electronics Engineering, IEEE, 2010.
- [10] FabrizioLamberti, Andrea Sanna, and Claudio Demartini, "A Relation-Based Page Rank Algorithm for Semantic Web Search Engines," IEEE Transactions on Knowledge and Data Engineering, VOL. 21, NO. 1, JANUARY, 2009.