A Hybrid Context based Approach for Web Information Retrieval

W. Aisha Banu
School of Decision and Information Sciences,
B.S.Abdur Rahman University,
Chennai, Tamil Nadu, India.

Dr. P. Sheikh Abdul Kader
School of Decision and Information Sciences,
B.S.Abdur Rahman University,
Chennai, Tamil Nadu, India.

ABSTRACT

Information retrieval mechanisms from the web are a great need of the hour as the amount of the content is growing dynamically every day. There are many algorithms which have been proposed in literature mainly relying on the output of the search engines. These algorithms are either content based or snippet based and perform a clustered outcome re-ranking of the content for the user. This work proposes a hybrid approach to content clustering that combines the best of the web information retrieval methods and also uses the personal preference information of the users modeling a wide range of contexts. This work introduces a context mechanism of the users in the overall process and presents taxonomy of the methods to organize the output of the search engines. Experimental results are promising and show that this approach has great promise for a wide range of queries.

Keywords

Web search, Context based search, Information retrieval

1. INTRODUCTION

The sheer volume of information growth in the web presents a combinatorial explosion of content that is quite unimaginable and yet daunting for a user. However, the users primarily use search engines for their content retrieval. In effect, the search engine itself is a combination of three distinct components [1]: a) the crawler and indexer which build the internal representation of the web for fast indexing; b) query database – this is the outcome of the crawling and indexing phase and represents a snapshot of the content in various forms and finally c) user query interface – here the query is processed according to relevance from the query database and the outcome is shown to the user.

In this, it is observed that the users rarely have the patience to navigate for content beyond the first five web result pages and look for other query terms. The broad outlines of search engine interaction are as follows: a) create query b) look at the results c) iterate d) modify query. In the recent years middleware systems have been developed that perform a bridge between the users query and the search engine. These middleware systems can intercede on behalf of the user with multiple search engines and cluster the content [2] based on probabilistic models. Supervised [3], Semi-Supervised [4] and unsupervised mechanisms [5] have been developed for the aggregation process.

Another development that is of great interest is the use of personalized search [6] based on the context of users and the use of context [7] of users. These systems create an internal snapshot of what the user wants and use query expansion techniques using ontologies to narrow down the scope of what the users want.

In applications like those for mobile users [8], the emphasis is on faster indexed retrieval rather than a phased information retrieval process. This work seeks to develop information retrieval mechanisms for the user contexts and hence focuses on maximum retrieval speed with mechanisms that can be built and modeled in the mobile devices. The ground work done by this paper can be later used in mobile information retrieval. Hence, while lessons from systems with large ontologies or multiple clustered outcomes from diverse sources or clustered tag based systems while more effective for a computer based user are taken into account, the focus is more on fast and hybrid information retrieval.

2. TAXONOMY OF WEB SNIPPET CLUSTERING

The taxonomy of the web snippet clustering algorithm (Figure 1) is given below. There are two different types of web snippet clustering techniques: a) Flat and b) Hierarchical. Flat web snippet clustering techniques do not consider the cluster relationships between the individual terms. In Hierarchical snippet clustering techniques, the relationships between the terms are considered. In both these techniques, we will either consider the phrases or consider the single sentences or take a collection of sentences. The techniques used for single terms in flat web snippet clustering are the Transactional K-Means clustering method or the Relational Fuzzy clustering. The techniques used in the sentences are the Suffix tree clustering and Singular Value Decomposition method. The frequent item set method is used in hierarchical technique. The lexical analysis, lexical affinities clustering, suffix array and topology based are the techniques used Hierarchical Web snippet Clustering method.

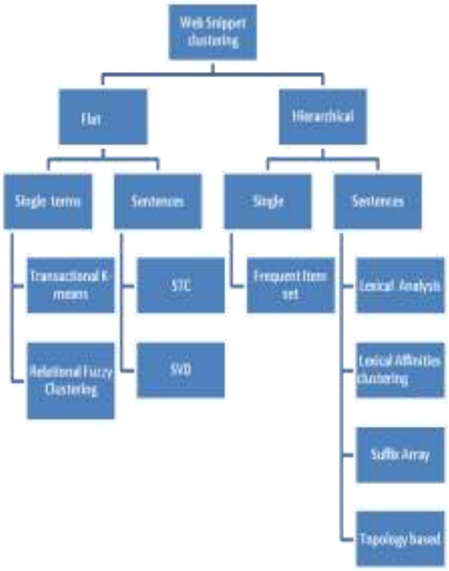

Figure 1: Web snippet clustering taxonomy

Our proposed work modifies the suffix tree clustering algorithm and uses a sentence based approach considering the relationship between the terms. Thus, it combines the best of the flat and hierarchical approaches in a hybrid manner for effective information retrieval.

3. HYBRID METHOD OF INFORMATION RETRIEVAL

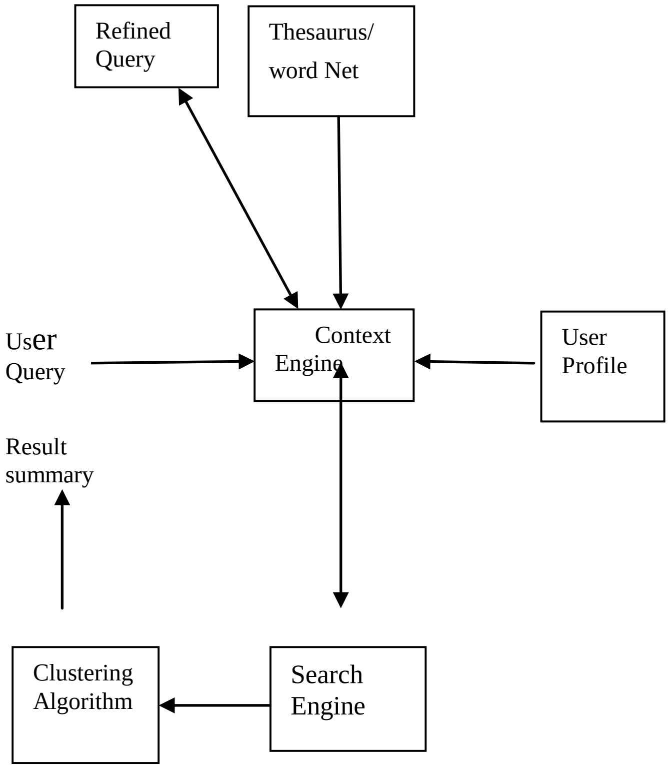

Figure 2: Proposed Hybrid Architecture

The proposed architecture (Figure 2) of the hybrid information retrieval process is shown below. The user's queries are processed by the context engine for the nature of query. Then based on the request type, the selected action is taken by the context engine.

The proposed work relies on the temporal nature of the search process of the mobile users. Hence the work combines the context information of the users and uses the snippet content weightage to guide the content re-ranking process. The context of the users is modeled in three ways a) content presentation b) content organization and c) content expansion.

The users can select to be in one of the modes of the search – fast, personal, expanded, balanced and relevant. These modes represent a combination of the three context properties. For a fast context, there is no expansion or context information used. The query is processed and immediately the outcome is served to the user. For a personal context, the query is compared against the database and the personal key words are served along with the query. The outcome of the search is a combination of the implicit personal preference and explicit query. In the expanded context, the query is expanded when compared to a database and the user is give additional pointers to narrow down the discourse of the query. Only when the user is certain about the content to navigate, the query is passed to the search engine. In the balanced context, along with the inputs from the personal preferences and the query expansion stages, the snippets of the outcome retrieved are modeled as a temporary stream and the comparison between query and the stream is found using snippet clustering mechanisms.

In a relevant context, the query outcomes (result pages) are combined into a common temporary stream. The query term and its related keywords are compared against this stream and result re-ranked according to the term modified document frequency. The snippet clustering mechanism, the query expansion system and the personal context information here serves as the elimination mechanism to reduce the domain of the discourse in an efficient manner. The primary focus of the algorithm is not just to find out how relevant the result is when compared to the query and overall set of query terms, but to cluster the results in an organized manner. This clustering system focuses on the content grouping and similarity indexes of documents. This mechanism is an unsupervised clustering system where a cluster is created when the document similarity exceeds the threshold (Log n) where n is the average number of words in the cluster.

The retrieval context hierarchy (Figure 3) is shown below.

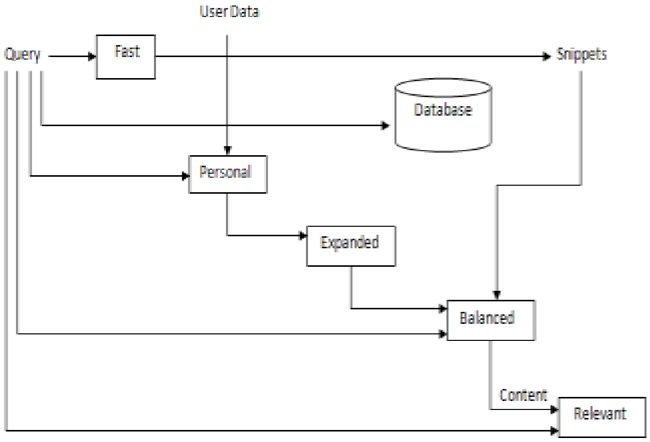

Figure 3: Proposed Context hierarchy

The hybrid clustering algorithm is shown below.

Hybrid Clustering algorithm

1: Extract all the n-grams as candidate words W = {w1, w2, ..., wi}, based on suffix tree built from a collection of web content D = {d1, d2, ..., dm}.
2: Build a word- document index R ={rp1 , rp2 , ..., rpl}, where rpk contains the indexed documents of Wk.

3: Construct a similarity measure model and calculate the similarity matrix for phrases using

$$\text{proximity(pi, pj)} = \boxed{= \frac{\left|\sum_{k=1}^{m} w_{ik} w_{jk}\right|}{\sqrt{\left(\sum_{k=1}^{m} w_{ik}^{2}\right)\left(\sum_{k=1}^{m} w_{jk}^{2}\right)}}}$$

4: While the number of clusters does not meet stopping criterion do

5: Merge the closest two clusters, and select a phrase of highest number of indexing documents from the merging clusters as the new cluster label.

6: Update the proximity matrix between the new cluster and the original clusters.

7: end while

8: Assign the contents whose indexing phrases belong to the same cluster.

9: Assign the remaining contents based on their k-nearest assigned neighbors.

Figure 4: Hybrid clustering algorithm

In this method, a modification of the suffix tree clustering algorithm [9] is used. In the fast suffix tree algorithm, there is exactly one node for each and every phrase in the document. A suffix tree has the feature by which a word can be inserted into the tree in a linear fashion by beginning from the root of the tree and using logic to decide whether to insert a new branch or split an existing branch into multiple branches. At each node we maintain the list of documents that contain the node and as well as an index that allows us to look up the phrase. The variation of the suffix tree algorithm is that we cross link the meanings of the words along with the words also. By doing so, multiple nodes in a document with the same meaning are clustered together as opposed to a linear fashion. The key to the overall process is the fact that the content stream is live and populated on the fly for a range of queries. Based on the outcome of the phase, a modified content tree is generated which takes into account the overall representation of the content. The overall outcome of this method is a set of clustered links and content along with the key content trees. The content trees now give the user an overview of what the snippets contain and is a concept map of the content.

The key to the overall process is that the context of users is temporal and periodic. Hence the past context of the users serves as a guideline for the current search. But at the same time, the user is provided with a range of options for processing.

4. EXPERIMENTAL RESULTS

The algorithm has been tested for a wide range of queries using a set of graduate students and a range of queries. Over 500 query results were used for comparison in a short duration of time. The key comparison was the relevance of the results and the time taken for the retrieval. For this, the baseline comparison was between the fast and other modes of retrieval. The experimental process consisted of the user noting down what exactly they were looking for before, during or after the query and which mode of retrieval suited them the best. Thus, for a query term “lightning”, the users intention was gauged as to whether it was for the term lightning as a natural occurrence or whether it was for a favorite team.

The first test metric (Figure 5) was the ability of the system to categorize content and generate new clusters of content effectively. For this a baseline comparison of the content was done and the performance of the algorithms projected

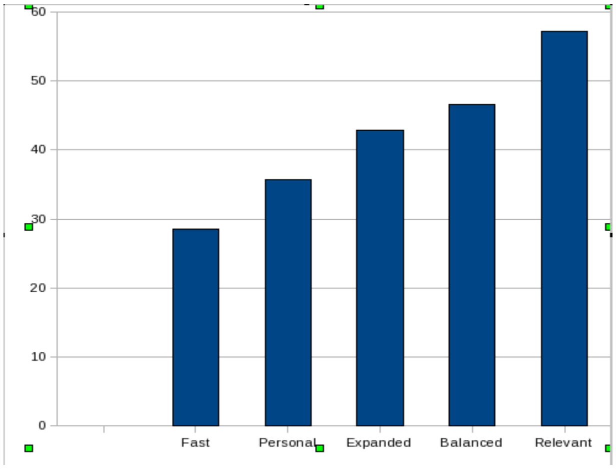

Figure 5: Cluster categorization efficiency

This shows that the fully content based retrieval system while not completely effectively in accurately predicting all the available clusters was able to accurately model the broad content categories more effectively than other modes of retrieval.

The second metric (Figure 6) is the external representation accuracy. The keys question here was if the content retrieved and shown represented an accurate perspective of what is actually contained in the document. For this, the output presented to the user and the actual documents were compared and the results tabulated.

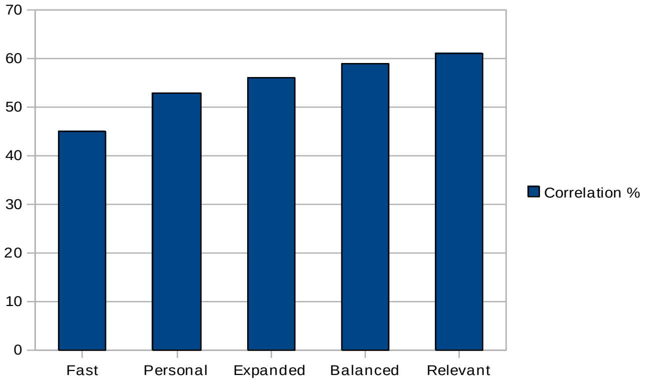

Figure 6: Correlation accuracy

Here the efficiency of the document mapping system is the key as the user gets not just the snippet tree, but also, the document based in a wide range of forms. Hence the outcome of the representation is a key to understanding what the users want.

The third metric (Figure 7) is the number of changes in the reranking system and the significance of the changes. For this, a composite criterion of ranking has been evolved. Thus a result reranking the third result to the first place is not a significant result. But a shift in the order from say the 20th result to the 9th result is considered significant.

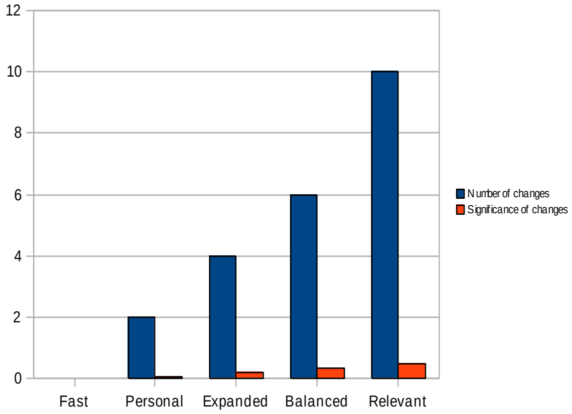

Figure 7: Significance of changes

The results show that there are a number of changes in the query output and the importance of the reranking has a small but distinct improvement in the output of the search display process. This improvement is not directly proportional to the amount of reordering or content changes, but improves the overall performance of the system by a distinct amount.

CONCLUSION

By this work, the need for fast context based information retrieval algorithms has been studied. The context behavior has been modeled in four distinct modes of operation and the contribution of each mode of operation to the overall information retrieval system has been quantified in terms of different parameters. These results show significant promise, but also underline the need for further work in the domain. In future, the algorithm will be implemented in a mobile context and the results tabulated. More testing in terms of precision and recall will be carried out with an expanded range of queries. The work will be compared with clustered search engines and algorithms in the future.

REFERENCES

[1] S. Nunes. 2007. Exploring temporal evidence in web information retrieval. BCS IRSG Symposium: Future Directions in Information Access.

[2] R. Manmat`ha, T. Rath and T. Feng. 2001. Modeling score distributions for combining the outputs of search engines. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York.

[3] Yu-Ting Liu, Tie-Yan Liu, Tao Qin, Zhi- Ming Ma, and Hang Li. 2007. Supervised rank aggregation. Analyzing Partially Ranked Data, In Proc. of the International World Wide Web Conference Notes in Statistics. Springer-Verlag, 1985. (WWW),

[4] L Si, J Callan. 2003. A Semi supervised Learning Method to Merge Search Engine Results , ACM Transactions on Information Systems (TOIS), - portal.acm.org.

[5] Alexandre Klementiev, Dan Roth, and Kevin Small. 2008. Unsupervised rank aggregation with distance-based models. In Proc. of the International Conference on Machine Learning (ICML).

[6] Chen, P.-M. and Kuo, F. 2000. C. An information retrieval system based on a user profile. J. Syst. Softw., 54 (1). 3-8.

[7] Mylonas, P. and Avrithis, Y., 2005. Context modeling for multimedia analysis. in Proc. of 5th International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT â€TM05), (Paris, France)

[8] icrosssing, How America searches Mobile, Technical Report. www.icrossing.com/.../How%20America%20Searches%20-%20Mobile.pdf

[9] Zamir, O., and Etzioni, O. 1998. Web document clustering: a feasibility demonstration. In SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, 46–54.New York, NY, USA: ACM