ISSN 2229-5518

Web Objects Clustering Through Aggregation for Enhanced Search Results

Dr. Pushpa R. Suri and Harmunish Taneja

Abstract— World Wide Web offer a rich mix of new challenges and opportunities to information computing researchers. The conventional search engine always returns a set of web pages in answer to a user query. Millions of web pages from organizations, institutions and personnel are made public electronically. With the web explosion and never ending raise of digital data, an added effect is the difficulty to retrieve relevant and reliable information from the Web. It is almost impossible for the naive user to get the right information in the answered search results as there is too much unrelated and out dated. The reason for this is rooted deep in the methodology for conventional Information computing on web that supports the indexing granularity for search as a web page. Search engines basically hunt for the potential web pages of user interests. On the contrary the user perspective in today's changing era is the information of a certain 'object' may be in the form of a cluster containing only the relevant data related to the object of interest rather than a tedious list of search results containing all the related and unrelated web pages. The similar theory can be applied to the queries from the point of view of developer. It requires grouping web objects into classes based on their attributes and links. This paper proposes algorithm for clustering web objects into different classes based on their links and identify relations dynamically. Results confirm the efficiency of the proposed approach as the user gets a cluster that contains only objects of interest from all the linked web pages.

Index Terms— Clustering, Information Retrieval (IR), Inheritance, Object Oriented, Search Engine, Web objects, World Wide Web (WWW).

1 INTRODUCTION

Exponential growth of the social networks on the WWW has result in an information explosion during recent years. Amount and variety of data is critical for the performance of the text retrieval systems. Information retrieval (IR) techniques evaluate user query and process complete series of potential web pages matching with some kinds of linguistic normalization of query keywords. Conventional IR expertise cannot be directly applied to collections of web pages due to the ever growing needs of the diverse users.

User does not want to be puzzled with the bulky set of web pages as search results of the said query. For example a query "vanilla ice cream" results in about eighty nine lacs web pages out of which all are not relevant and it is highly unlikely that a user interested in vanilla ice cream will open all search results not to mention the enormous wastage of the extraction and indexing effort in computing such vast information. The reason for this lapse is very clear; the basic search quantum is a web page and the user need to search an object "vanilla ice cream".

This gap can be bridged by tailoring the web page to the cluster of objects of interest and this proposed concept can be extended to aggregating the related potential desired web objects in to cluster and perform information computing on it. Web search engines inherit many disadvantages of traditional information retrieval Systems [1]. Simple query representation scheme where a web page is typically expected by a list of keywords or a Boolean expression totally overrules the relationships between different web pages and between different parts of the same web page. This paper suggests a two tier definition of the web objects. One is the user objects submitted to the WWW in the form of queries. The other one is the objects users takes from the web as the search results. Web objects are defined to represent significant objects embedded in web pages and pointed to by the hyperlinks.The proposed approach indicates a modern way to use web knowledge to help organize vastly distributive information of facade web in to more structured clusters of related objects. For cluster identification it is noted that complementary information of the web object is highly likely to be present in the web pages that are hyperlinked [15]. For example, users may use query "researcher Tim Burners-Lee " to search for the biography about Tim Burners-Lee, Father of web, his books and publications, instead of the web pages that contain the query terms only. To meet this contemporary search need, an 'object' oriented approach to information computing is obligatory. The major benefit of proposed clustering is two folded due to its basic behaviour of containing similar objects. If one object is irrelevant from the user perspective, then the whole cluster can be eliminated thereby reducing the complexity of the computing by the order of the cluster

Dr. Pushpa R. Suri is working as Associate Professor in the department of Computer Science and Applications at Kurukshetra University Kurukshetra. pushpa.suri@yahoo.com
Harmunish Taneja is a Research Scholar at Kurukshetra University Kurukshetra and working as Asstt. Prof. in the Department of Information Technology, Maharashi Markandeshwar University, Mullana, Haryana India. harmunish.taneja@gmail.com

size. The same is true for selecting a relevant cluster. The rest of this paper is organized as follows: Section 2 presents the related work. Section 3 elaborates the proposed clustering framework. Finally, the proposed approach is evaluated in Section 4 and conclusions are given in Section 5.

2. Related Work

The applications of web objects clustering are to cluster web objects into categories so as to facilitate the information computing on web in terms of information retrieval and filtering [4]. Traditional clustering algorithms [2] divide objects into classes where object in a class are similar to each other and dissimilar from objects in different classes. IR based on clustering web objects is different from traditional clustering problem as it categorizes the objects into the classes on the basis of their properties as well as links. The major challenge of desired clustering analysis is the object division in the classes based on objects links [15]. Conventional clustering methods define similarities among different objects by distances over content features of objects [6]. A web document is often represented as a vector in a vector space formed by all the keywords [5]. Clustering objects from the user perspective is harder as queries lacks in features extraction directly as they are of few keywords only [3]. Traditional clustering methods ignore the link information and clusters object solely based on content features [9][10][11]. To discover related objects hyperlinks and click-through data are considered in Google's PageRank algorithm weighting hyperlinks [7]. Bipartite graph between queries and web pages is created for clustering [8]. User clustering method may also be based on access patterns, where the access patterns are generalized like web-pages with the same URL prefix are in same cluster [12]. Query clustering method using user logs is suggested that quote two queries as similar provided they contain similar terms or result in the selection of similar web pages as search results [13]. A correlation-based document clustering method measures the similarity between web-pages based on the simultaneous visits to them [14].

3. Clustering Framework for the Web Objects

Proposed clustering framework reorganizes query results by exploiting the link analysis with the object-oriented approach to identify similar web objects. The feature vector is defined in the combination to the object oriented database and object lexicons created for each domain. The work distinguishes itself from existing work by incorporating inheritance network tree construction for the query words entered by the user.

Information retrieval systems are everywhere from cook book indexes to library catalogs and last but not the least web search engines. The web information retrieval system generates the set of the search results in the form of web pages corresponding to the user's query as shown in working of traditional IR in Fig. 1. IR comprises of components like Indexer, ranker, sorter and Crawler. Crawler runs through the web and gathers the required information for the search engine. The Indexing process collects, parses and stores data in order to facilitate fast and accurate IR. Ranker evaluates the rank of the web

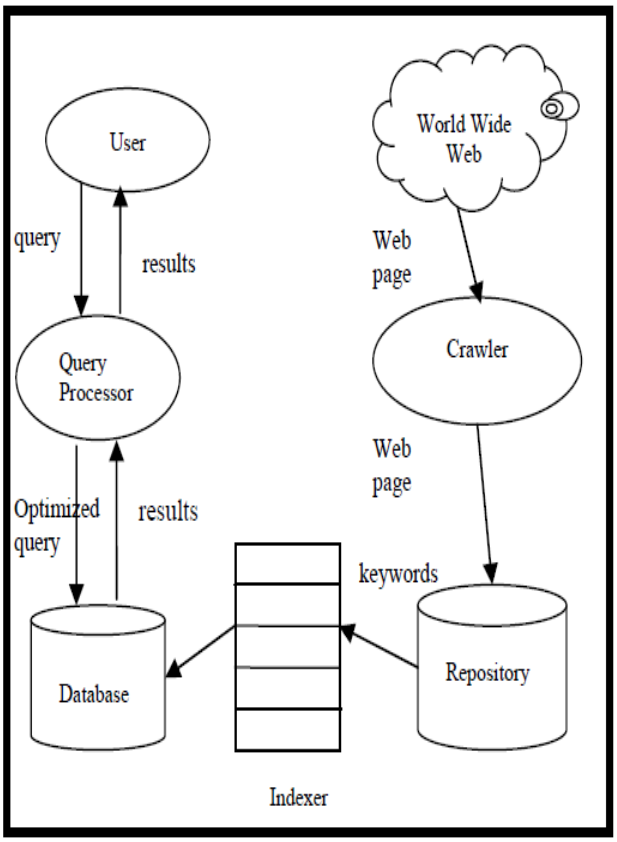

Fig. 1: General architecture of the Search Engine [16]

pages based on number of hits and links. The proposed work as shown in the Fig. 2 is the expansion of the architect ture given in Fig. 1. Unlike Fig. 1, the architecture proposed in Fig. 2 does not return the retrived search directly to the user. Instead it is further feeded to the inheritance induced relation recognizing unit where a threshold for the feature similarity decides the selection and rejection of the web objects in the clusters. The automatic creation of inheritance network tree assigns the freature similarity value to each web object based on the relation network graph. Each object is then accepted or rejected for the membership to the cluster based on the feature similarity value check against a pre defined constant delta. The algorithm Web_Object_Inheritance_Cluster algorithm for the same is elaborated in table 1.1. The clusters represent strong or loose coupling among the member web objects as shown in Fig. 3.

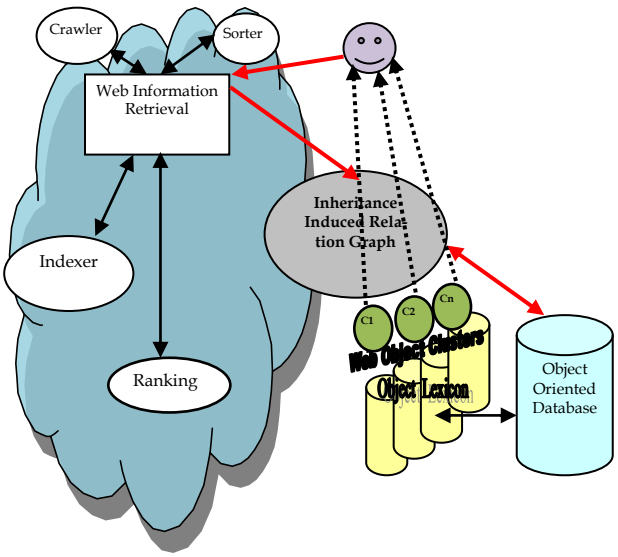

Fig. 2: Clustering framework for enhanced IR

Table 1.1. Algorithm: Web_Object_Inheritance_Cluster
Input: Web objects occurring corresponding to query Q.
Output: Cluster of size n of related web objects
Begin Step 1: User send the keyword based query to the web IR system. Step 2: Vertical indexed search results as web pages are retrieved and sent to the inheritance induced relation recognizer. Step 3: Relation recognizer constructs the inheritance tree with all the related objects as leaf of a common parent at distinct level based on the assigned feature similarity values. Step 4: IF{Web object *feature similarity* F_s < *delta*} THEN object accepted for cluster membership ELSE repeat step 3 for all objects F_s where s < n Step 5: Cluster with related web object corresponding to the query is sent to the user End.

As shown in Fig. 3, the inheritance induced relation graph is generated after all the candidate web objects are searched for possible links among them with the help of object lexicon and object oriented database. O1, O2, O3 inherit attributes from a general class and hence the entire web objects that may belong to one or ore subclasses are aggregated as a cluster. O11, O12 are derived from object O1. Similarly O311 and O322 share link with the common object O31 which is further inherited from O3 and so on. Such an inheritance based graph assures that each cluster contains only relevant and reliable information as all the member objects passes the feature similarity above a threshold.

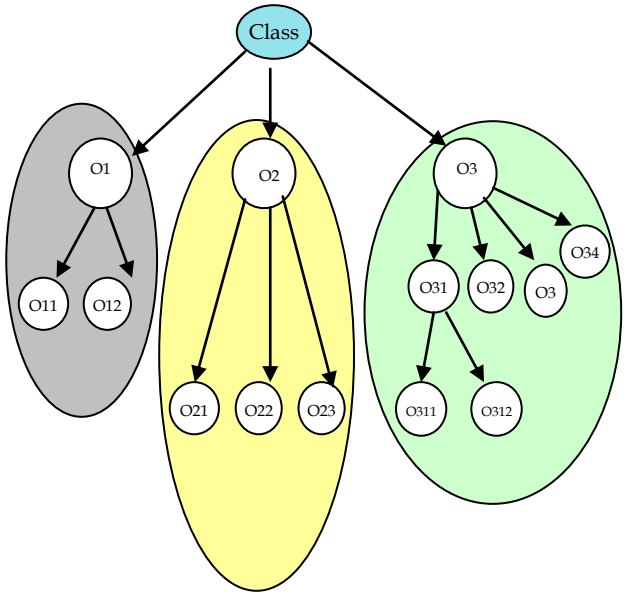

Fig. 3: Inheritance induced Clusters

4. Experimentation

The text based web search queries are performed on several large search engines and collecting the first 50 results for each. Queries used are framed to give large number of similar kind of web documents. Queries like publications, research scholars, biography are chosen. For non duplicate data thousands of web objects are chosen at random. As the feature threshold is varied, all the web objects which satisfies the threshold of one cluster, becomes its member. The Web_Object_Inheritance_Cluster algorithm proves to be fast for the small set. The difference in speed will increase with the size of the cluster. As the size and number of clusters per user query increases, the overhead increases but still levels off for medium sized datasets and the performance is increased due to fast convergence. The reason is few candidate objects of the cluster are checked and if found irrelevant the entire cluster is eliminated from future information computing for that particular query.

5. Conclusion

Web object represents significant object embedded in web pages or linked to by hyperlinks. Users usually search for information of a certain 'object', rather than a large set of web pages resulting from a keyword based query. An algorithm for building clusters of related web objects distributed over the web which can be used for fast user friendly online search is proposed. The work supports automatic creation of inheritance induced relatedness tree in the directed graph form that ssigns the feature similarity value to each web object which is accepted / rejected from the membership of the cluster if the feature vector threshold value is satisfied. The threshold value is a constant and uniformly assumed in advance and may not be

appropriate for heterogeneous domains. Algorithm is independent of the object type or domain and may be extended to domain specific feature similarity vector computation.

REFERENCES

[1] S. Lawrence, and G.L. Giles, "Context and Page Analysis for Improved Web search", *IEEE Internet Computing*, vol. 2, pp. 38-46, 1998. (Journal citation)

[2] Han J., and Kamber M., *Data Mining: Concepts and Techniques*, 2nd Ed., San Francisco: The Morgan Kaufmann Publishers, 2006. (Book style)

[3] Jia Rongfei, Jin Maozhong, and Wang Xiaobo, "Web Objects Clustering Using Transaction Log," *Proc. Third International Conference on Knowledge Discovery and Data Mining by IEEE Computer Society*, pp. 182-186, 2010. (Conference Peoceedings)

[4] W. Chan, W. Leung, and D. Lee, "Clustering Search Engine Query Log Containing Noisy Clickthroughs," *Proc. SAINT Conference*, Tokyo, Japan, pp. 305-308, 2004. (Conference Peoceedings)

[5] R. Xu, and I. Donald Wunsch, "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645, 2005. (IEEE Transactions)

[6] A. Strehl, J. Ghosh, and R. Mooney, "Impact of Similarity Measures on Web-page Clustering," *Proc. AAAI Workshop on AI for Web Search (AAAI 2000)*, Austin, pp. 58–64, 2000 (Conference Peoceedings)

[7] S. Brin and L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998. (Journal citation)

[8] D. Beeferman, and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," *Proc. of the Sixth ACM SIGKDD International conference on Knowledge discovery and data mining*, ACM New York, NY, USA, pp. 407–416, 2000. (Conference Peoceedings)

[9] D. Boley, M. Gini, R. Gross, E.H. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore, "Partitioning-based Clustering for Web Document Categorization," *Decision Support Systems,* vol. 27, pp. 329-341, 1999. (Journal citation)

[10] O. Zamir and O. Etzioni, "Web Document Clustering: A Feasibility Demonstration," *Proc. of SIGIR '98*, pp. 46--53, 1998. (Conference Peoceedings)

[11] A. Strehl, J. Ghosh, and R. Mooney, "Impact of Similarity Measures on Web-page Clustering," *Proc. of the AAAI 2000 Workshop on Artificial Intelligence for WebSearch*, pp. 58--64, Austin, Texas, July 2000. (Conference Peoceedings)

[12] Y. Fu, K. Sandhu, and M. Shih, "Clustering of Web Users Based on Access Patterns," *Proc. of the 1999 KDD Workshop on Web Mining*, San Diego, Canada, 1999. (Conference Peoceedings)

[13] J. Wen, J.Y. Nie, H. Zhang, "Query Clustering Using User Logs," *ACM Transactions on Information Systems*, vol. 20, no. 1, pp. 59-81, 2002. (ACM Transactions)

[14] Z. Su, Q. Yang, H. J. Zhang, X. Xu and Y. H. Hu, "Correlation-based Document Clustering using Web Logs," *Proc. of the 34th Hawaii International Conference On System Sciences (HICSS-34)*, January 3-6, pp. 1-7, 2001. (Conference Peoceedings)

[15] Sanjeev Sharma, R.K. Gupta, "Improved BSP Clustering Algorithm For Social Network Analysis," *International Journal of Grid and Distributed Computing*, vol. 2, no. 3, pp. 67-76, Sept. 2010. (Journal citation)

[16] Deepti Gupta, Komal Kumar Bhatia, and A.K. Sharma, "A Novel Indexing Technique for Web Documents using Hierarchical Clustering," *IJCSNS International Journal of Computer Science and Network Security*, vol.9 no.9, pp. 168-175, , September 2009. (Journal citation)

Dr. Pushpa R. Suri received her Ph.D. Degree from Kurukshetra University, Kurukshetra. She is working as Associate Professor in the Department of Computer Science and Applications at Kurukshetra University, Kurukshetra, Haryana, India. She has many publications in International and National Journals and Conferences. Her teaching and research activities include Discrete Mathematical Structure, Data Structures, Information Computing and Database Systems.

Harmunish Taneja received his M.Phil. degree in (Computer Science) from Algappa University, Tamil Nadu, India and Master of Computer Applications from Guru Jambeshwar University of Science and Technology, Hissar, Haryana, India. Presently he is working as Assistant Professor in Information Technology Department of M.M. University, Mullana, Haryana, India. He is pursuing Ph.D. (Computer Science) from Kurukshetra University, Kurukshetra. He has published 14 papers in International / National Conferences and Seminars. His teaching and research areas include Database systems, Web Information Retrieval, and Object Oriented Information Computing.