ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (831-834), Month: April - June 2015, Available at: www.researchpublish.com

Ant Based Web Content Clustering Based On Personalization of Users

[1]Preethi.C, [2]K.Bhuvaneswari, [3]G.Sudhakar

[1, 2] PG Scholar, Department of Computer Science and Engineering, Ranganathan Engineering College, Coimbatore, India
[3]Head of the Department, Department of Computer Science and Engineering, Ranganathan Engineering College, Coimbatore, India

***Abstract:* Ant-Based Web Content Clustering based on personalization of users uses Ant-based Clustering algorithm for efficient clustering of web contents and provide the users with the best solution. Query log analysis has emerged as one of the most promising research areas to automatically derive such structures. A biologically inspired model based on Ant Colony Optimization applied to query logs as an adaptive learning process that addresses the problem of deriving query suggestions is explored. A user interacts with the ranking system at pre-computation and query times. During pre-computation, the user-specific personalized Weight Assignment Vector can be computed using (a) It can done by domain experts according to the user's profile (b) It can be learned automatically by user interaction and exploiting user relevance feedback User interactions with the search engine which is treated as an individual ant's journey and over time the collective journeys of all ants result in strengthening more popular paths which leads to a corresponding term association graph that is used to provide query modification suggestion which is updated in a continuous learning cycle. Ant Colony based Clustering has been studied extensively as a form of swarm intelligence technique to solve problems in several domains such as Scheduling, Classification and Routing problems.**

***Keywords:* personalization, Ant colony, WAV.**

I. INTRODUCTION

In today's information system management, large-scale data clustering and classification have become increasingly important and a challenging area. Although various tools and methods have been proposed, few are sufficient and efficient enough for real applications due to the exponential growing-in-size and high-dimensional data inputs. As a significant application area of data mining is Web Mining (WM). Web mining is the application of data mining techniques to extract knowledge from web data, i.e. web content, web structure, and web usage data. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.

Web usage mining itself can be classified further depending on the kind of usage data considered as Web Server, Application Server and Application Level data.

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and Computer vision. Data mining is the main concept used in retrieving the data from web pages.

II. RELATED WORK

Dynamic Personalized Page ranking Entity-Relation Graphs[3] presents HubRank, a new system for fast, dynamic, space efficient proximity searches in ER graphs. During preprocessing, HubRank computes and indexes certain “sketchy” random walk fingerprints for a small fraction of nodes, carefully chosen using query log statistics. We present HubRank, a new system for fast, dynamic, space efficient proximity searches in ER graphs. During preprocessing, HubRank computes and indexes certain “sketchy” random walk fingerprints for a small fraction of nodes, carefully chosen using query log statistics. At query time, a small “active” sub graph is identified, bordered by nodes with indexed fingerprints. These fingerprints are adaptively loaded to various resolutions to form approximate personalized Pagerank vectors (PPVs). PPVs at remaining active nodes are now computed iteratively. We report on experiments with CiteSeer’s ER graph and millions of real Cite- Seer queries. Scalable Link-based Personalization for Ranking in Entity-Relationship Graphs[4] proposes for the first time an approach to achieve efficient edge-based personalization using a combination of pre-computation and runtime algorithms. Authority flow techniques like PageRank and ObjectRank can provide personalized ranking of typed entity- Relationship graphs. There are two main ways to personalize authority flow ranking: Node based personalization, where authority originates from a set of user specific nodes; Edge-based personalization, where the importance of different edge types is user-specific. We propose for the first time an approach to achieve efficient edge-based personalization using a combination of pre-computation and runtime algorithms. We apply our method to the personalized authority flow bounds of ObjectRank, i.e., Weight Assignment Vector (WAV) assigns different weights to each edge type or relationship type. Our approach includes a repository of rankings for various WAVs. We consider the following two classes of approximation: (a) SchemaApprox is formulated as a distance minimization problem at the schema level; (b) DataApprox is a distance minimization problem at the data graph level. SchemaApprox is not robust since it does not distinguish between important and trivial edge types based on the edge distribution in the data graph. Both SchemaApprox and DataApprox are expensive so we develop efficient heuristic implementations. ScaleRank is an efficient linear programming solution to DataApprox. PickOne is a greedy heuristic for SchemaApprox. Extensive experiments on the DBLP data graph show that ScaleRank provides a fast and accurate personalized authority flow ranking. We propose two heuristics, ScaleRank and PickOne, as DataApprox and SchemaApprox respectively are too expensive to facilitate interactive query reponse. Authoritative Sources in a Hyperlinked Environment [5] presented an adjustable framework to answer keyword queries using the authority transfer paradigm, which we believe is applicable to a significant number of domains (though obviously not meaningful for every database). We showed that our framework is efficient and semantically meaningful, with an experimental evaluation and user surveys respectively. It propose and test an algorithmic formulation of the notion of authority, based on the relationship between a set of relevant authoritative pages and the set of “hub pages” that join them together in the link structure. We first present how state-of-the-art works rank the results of a keyword query, using traditional IR techniques and exploiting the link structure of the data graph. Then we discuss about related work on the performance of link-based algorithms. The user survey investigates and compares alternative ways to incorporate link-based specificity to keyword queries. In particular, we propose alternative specificity metrics and also experiment with various ways to incorporate Inverse ObjectRank in the ranking. Authority based keyword search in databases[6] provides technique for locating high-quality information related to a broad search topic on the www, based on a structural analysis of the link topology surrounding “authoritative” pages on the topic. We propose and test an algorithmic formulation of the notion of authority, based on the relationship between a set of relevant authoritative pages and the set of hub pages" that join them together in the link structure. We develop a set of algorithmic tools for extracting information from the link structures of such environments, and report on experiments that demonstrate their effectiveness in a variety of contexts on the World Wide Web. The www is a hypertext corpus of enormous complexity, and it continues to expand at a phenomenal rate. Moreover, it can be viewed as an intricate form of populist hypermedia, in which millions of on-line participants, with diverse and often conflicting goals, are continuously creating hyperlinked content. We begin from the observation that improving the quality of search methods on the www is, at the present time, a rich and interesting problem that is in many ways orthogonal to concerns of algorithmic efficiency and storage. The results are of as a high quality as possible in the context of what is available on the www globally. BinRank: Scaling Dynamic Authority-Based Search Using Materialized Subgraphs[10] introduces BinRank, a system that approximates ObjectRank results by utilizing a hybrid approach inspired by materialized views in traditional query processing. We proposed BinRank as a practical solution for scalable dynamic authority-based ranking. It is based on partitioning and approximation using a number of materialized subgraphs. PPR is a modification of PageRank that performs search personalized on a preference set that contains Web pages that a user likes. For a given preference set, PPR performs a very expensive fixpoint iterative computation over the entire Web graph, while it generates personalized search results. In this paper, we introduce a BinRank system that employs a hybrid approach where query time can be traded off for preprocessing time and storage. BinRank closely approximates ObjectRank scores by running the same ObjectRank algorithm on a small subgraph, instead of the full data graph. The subgraphs are precomputed offline. The precomputation can be parallelized with linear scalability. Object-Level Ranking: Bringing Order to Web Objects calculates the object popularity scores of Web objects based on their Web popularity and the object relationship graph. This paper introduces PopRank, a domain-independent object-level link analysis model to rank the objects within a specific domain. Specifically we assign a popularity propagation factor to each type of object relationship, study how different popularity propagation factors for these heterogeneous relationships could affect the popularity ranking, and propose efficient approaches to automatically decide these factors. PopRank extends the PageRank model by adding a popularity propagation factor (PPF) to each link pointing to an object, and uses different propagation factors for links of different types of relationships. We propose a learning based approach to automatically learn the popularity propagation factors for different types of links using the partial ranking of the objects given by domain experts. The simulated annealing algorithm is used to explore the search space of all possible combinations of propagation factors and to iteratively reduce the difference between the partial ranking from the domain experts and that from our learned model. It automatically assigns a popularity propagation factor for each type of object Relationship. PopRank can achieve significantly better ranking results than naively applying PageRank on the object graph. This method did not evaluate the effectiveness of the ranking scheme when Web pages are intermixed with semantic objects. It currently works on evaluating the model in a more general way and in other application domains.

III. ANT COLONY ALGORITHM

Ant Based Clustering Method is used to cluster the data in an efficient manner and to retrieve the best results within a short span of time and also with great accuracy when compared to the ScaleRank algorithm. Here the web pages are considered as ants. Query is given by the user, in which the algorithm sees for Matching, Neighborhood and Distance time parameters to find the web page that best matches the query. In the first step, it finds all the web pages. In second step, it takes the best pages based on the above mentioned parameters and it updates the pheromone accordingly. The pheromone indicates the best path to reach the page. Here no needs for top K algorithm, as the best results are obtained early. The system can handle large volumes of data, No need of top K algorithm, Has greater accuracy and takes less time than ScaleRank algorithm and it also performs pheromone updation automatically even if a best path comes later.

The Ant based web content clustering and retrieval is obtained by the following modules:

- User Interaction or Profile Identification
- Comparing with Personalized Repository
- Ant Colony Algorithm Implementation
- Query results and Repository Updation

3.1 User Interaction or Profile Identification:

This module is for getting the query from the user. A GUI application is created to get the user details for registration and the Search query. Every user needs to register in that GUI application with their details. And if the user gives the query in the search, each query is noted with the particular search time and it is stored in the database. The current user is identified by using session attributes in jsp. It finds out the user preferences or personalization of the users.

3.2 Comparing with Personalized Repository:

To identify the relevant queries, we need to perform query reformulations that are typically found within the query logs of a search engine. If two queries that are issued consecutively by many users occur frequently enough, they are likely to be reformulations of each other. To measure the relevance between two queries issued by a user, the time based metric, sim time, makes use of the interval between the timestamps of the queries within the user's search history. In contrast, our approach is defined by the statistical frequency with which two queries appear next to each other in the entire query log, over all of the users of the system.

3.3 Ant Colony Algorithm Implementation:

The aim of data-clustering is to obtain optimal assignment of N objects in one of the K clusters where N is the number object popularity scores of Web objects based on their Web popularity and the object relationship graph .of objects and K is the number of clusters. In Clustering with Ant Colony Optimization, ants start with empty solution strings and in the first iteration the elements of the pheromone matrix are initialized to the same values. With the progress of iterations, the pheromone matrix is updated depending upon the quality of solutions produced. Each ant selects a cluster number with a probability value for each element of S string to form its own solution string S. The quality of constructed solution string S is measured in terms of the value of objective function for a given data-clustering problem. This objective function is defined as the sum of squared Euclidian distances between each object and the center of belonging cluster. Then, the elements of the population, namely agents are sorted increasingly by the objective function values. Because, the lower objective function value, the higher fitness to the real solution, namely, lower objective function values are more approximated to real solution values. An optimal solution is that solution which minimizes the objective function value. This process explains that an iteration of the algorithm is finished. Algorithm iterates these steps repeatedly until a certain number of iterations and solution having lowest function value represents the optimal partitioning of objects of a given dataset into several groups.

3.4 Query results and Repository Updation:

Each query group contains closely related and relevant queries and clicks, it is important to have a suitable relevance measure between the current query singleton group sc and an existing query group si. There are a number of possible approaches to determine the relevance between sc and si. A relevance measure that is robust enough to identify similar query groups beyond the approaches that simply rely on the textual content of queries or time interval between them. Our approach makes use of search logs in order to determine the relevance between query groups more effectively. Finally, the results are analyzed by comparing the execution time and accuracy of our system with the ScaleRank approach.

IV. CONCLUSION AND FUTURE WORK

In this work, we have presented Ant-based clustering method which clusters and sorts data in an efficient manner. They perform both vector quantization and topographic mapping at the same time. They have linear scaling behavior. They can handle clusters of any shape. This method takes less processing time to identify the user preferences and also groups all web contents accurately as per user needs and present them to the user based on the query given. Future developments of this work are related further to improve our mathematical model by adding more Quality of Service parameters. In the future, we plan to further improve our algorithm to efficiently perform the search and provide the user with the best solution at a minimum cost and response time.

REFERENCES

[1] Vagelis Hristidis, Yao Wu, and Louiqa Raschid,” Efficient Ranking on Entitiy Graphs with Personalized Relationships”, IEEE, Transactions on Knowledge and Data Engineering, Vol No. 4, April 2014.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank citation ranking: Bringing order to the web,” Stanford Digital Library Technologies Project, Tech. Rep., 1998.

[3] S. Chakrabarti, “Dynamic personalized PageRank in entity-relation graphs,” in WWW ’07. New York, NY, USA: ACM, 2007, pp. 571–580.

[4] V.Hristidis, L.Raschid, and Y.Wu, “Scalable Link-Based Personalization for Ranking in Entity-Relationship Graphs”, in SIGMOD, 2011, pp. 552– 563.

[5] J.M.Kleinberg, “Authoritative Sources in a Hyperlinked Environment”, J.ACM Trans. vol. 46, no. 5, pp. 604–632, 2008.

[6] V.Hristidis, H.Hwang, and Y.Papakonstantinou, “Authoritative Based keyword Search in Databases”, in ACM, Vol.3, pp. 1-40,2008.

[7] A. Balmin, V. Hristidis, and Y. Papakonstantinou, “Objectrank: Authority-based keyword search in databases.” in VLDB, 2004, pp. 564–575.

[8] D. Fogaras, B. R ´ acz, K. Csalog´ any, and T. Sarl ´ os, “Towards scaling fully personalized PageRank: Algorithms, lower bounds, and experiments,” Internet Mathematics, vol. 2, no. 3, 2005.

[9] T. H. Haveliwala, “Topic-sensitive PageRank,” in WWW ’02.

[10] H. Hwang, A. Balmin, B. Reinwald, and E. Nijkamp, “BinRank: Scaling dynamic authority-based search using materialized subgraphs,” in ICDE ’09, 2009, pp. 66–77.

[11] G. Pandurangan, P. Raghavan, and E. Upfal, “Using PageRank to Characterize Web Structure,” Computing and Combinatorics, O. Ibarra, and L. Zhang, eds., pp. 1-4, 2002.