Web Personalization and Mining Using Robust Fuzzy Clustering Methods

Affiliation & Contact Information

Anupam Joshi
Dept. of Computer Science and Electrical Engg.

University of Maryland, Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250
Phone: (410) 455-2590
Fax : (410) 455-3969
Email: joshi@cs.umbc.edu

Raghu Krishnapuram
Dept. of Mathematical and Computer Sciences

Colorado School of Mines
1005 West 14th Street
Golden, CO 80401
Phone: (303) 273-3877
Fax : (303) 273-3875
Email: rkrishna@mines.edu

WWW Page: http://www.cs.umbc.edu/~joshi/web-mine/

List of Supported Students and Staff

Liyu Yi, CSM, GRA (9/98 - 12/99), Scott Jiang, UMBC, GRA (1/99 -05/00), Kejian Hu, UMBC, GRA (1/2000 - 5/2000), Tapan Kamdar, GRA, UMBC (9/1999 -), Olfa Nasraoui, CSM, GRA (3/99 - 7/99)

Project Award Information

- Award Numbers: IIS 9801711 and IIS 9800899
- Duration: 9/01/1998 - 8/31/2001, Presently in Year 3
- Title: Web Personalization and Mining Using Robust Fuzzy Clustering Methods

Keywords

Web Mining, Information Personalization, Robust Fuzzy Clustering, Adaptive Web Sites, Web User Categorization, Web Document Classification, Snippet Analysis, Web Access Patterns, Clickstreams

Project Summary

This inter-institutional collaborative project is being carried out by Anupam Joshi at the University of Maryland, Baltimore County, and Raghu Krishnapuram at the Colorado School of Mines. The goal is to develop scalable robust techniques to model noisy data sets containing an unknown number of overlapping categories, and apply them to create a software tool for Web personalization and mining. Personalization has two components: (1) tailoring the content delivered to the user from a Web site; and (2) exploring the available Web pages and categorizing them. The approach consists of developing new practical clustering algorithms by combining fuzzy methods and robust statistics to model an unknown number of overlapping sets in the presence outliers. Since many Web objects such as URLs, IP addresses, Web pages, and snippets cannot be represented by numerical features, new techniques to handle such Web objects, as well as to categorize them by using suitable similarity measures between such objects, are being explored. A software tool for Web personalization and mining is being created and validated. The results of this project will generate new theoretical results and efficient algorithms for simultaneously estimating the prototypes/profiles of an unknown number of overlapping categories from noisy data sets, as well as a Web personalization and mining tool that will be made available on the Web. Thus, this project is expected to have a significant impact on the way documents are searched for and delivered. It will directly influence the usefulness and spread of the WWW, and will also contribute to the digital library technology.

Goals, Objectives, and Targeted Activities

Our third year goals are:

1. To continue development and refining of new robust fuzzy relational clustering algorithms that are suitable for large applications, and test them on real data sets. In particular, we are creating an incremental version of our clustering algorithm.
2. To analyze the differences between browsing sessions created with and without the use of cookies.
3. To create prototype systems for Web Log Mining and Web Search Engine Results Clustering.

Indication of Success

Our project started on September 1, 1998, and is in its third year. Our work has lead to the development of several new fuzzy clustering algorithms that are robust in terms of noise handling. We have also shown that while the overall average complexity of these algorithms is O(n^2), we can create an O(n) version that performs just as well. We are presently working on an incremental version of the O(n) algorithm that will make it even more useful in creating adaptive web sites, and are on target to finish this over the summer. We have created two application system prototypes that use these algorithms.

The first system analyzes Web logs to recover the "traversal path" of the user through the server. We have defined a non-Euclidean similarity measure between paths, and used it in two new robust fuzzy relational clustering algorithms (Genetic Niche Clustering and Trimmed Fuzzy C-Medoids), we are able to find "typical" traversal paths through Web servers at two different universities and other organizations. This information can be useful in two ways. One, it can let the site developer know about the associations between the various URLs in their site. Secondly, it can be used to make the Web site adaptive in conjunction with state-maintaining devices such as cookies. In the past year and a half , we have shown that the algorithms can be made to scale linearly, and have conducted several experiments to better characterize their noise handling ability. We have also created a prototype adaptive web site system that creates different views of the same web site for different users. This system operates on the UMBC CS web site with a separate development server. The algorithm thus gets tested on a fairly large and fairly active (30-40K hits/day) web site.

The second system applies the developed algorithms to the task of grouping URLs returned by a search engine on the fly. Most search engines simply return a large list of URLs in response to a query. The idea here is to use the snippets associated with the returned URLs to see if these URLs can be grouped meaningfully, with keyword descriptions for each group. This would enable a user to focus only on a relevant subset of URLs based on the associated keywords. Such grouping has been explored in the IR community for text documents, but doing it for Web search engines presents new problems relating to noise and scalability. Given that we group based only on the snippets, the SNR of the data is fairly low. Moreover, this grouping needs to be done in near real time to be acceptable to users.

Project Impact

We have created new algorithms as well as prototype systems for two important problems -- Web log mining and search engine result clustering. Several papers describing the work have been published. The prototypes will be made publically accessible in the next few months. We will also present this work to colleagues in industrial labs. A total of four students (one Ph.D and three master's) supported by this grant at Mines and UMBC have already graduated and are employed in the industry. Two more students will graduate over the summer of 2001 at UMBC. We note that this past year, Krishnapuram's grant (9800899) was on hold as he was on a leave of absence from Mines. Joshi taught a course in the spring of 1999 on Web and Data Mining. The course was heavily oversubscribed.

GPRA Outcome Goals

1. Discoveries at and across the frontier of science and engineering:
   A brief discussion of discoveries to date is discussed under "Indications of Success" above.
2. Connections between discoveries and their use in service to society:
   The research will lead to Web sites whose content is tailored automatically to suit the interests of the user. The work on search engine results clustering will will not only lead to a better, more personalized experience for Web users, but also cut down on the network traffic generated as users will not have to sift through needless information.

3. A diverse, globally oriented workforce of scientists and engineers:
   The graduate students working on this project, and the undergraduate students in the classes being taught by the PIs in this are diverse and being trained in this area of increasing importance to the Global Information Infrastructure.

Project References

1. Robust Fuzzy Clustering Methods to Support Web Mining, Proc. ACM SIGMOD Workshop on Data Mining and Knowledge Dis covery, Aug 98.
2. Personalization & Asynchronicity to Support Mobile Web Access, *Proc. Workshop on Web Information and Data Management, ACM Conf. on Information and Knowledge Management*, Bethesda, MD, Nov. 1998.
3. Mining Web Access Logs Using Relational Competitive Fuzzy Clustering, Proc. *Eight International Fuzzy Systems Association World Congress - IFSA 99*, August 1999.
4. Mining Web Access Logs Using a Fuzzy Relational Clustering Algorithm based on a Robust Estimator, (Poster) Proc. *8th World Wide Web Conference (WWW8)*, Torronto, May 1999.
5. A Fuzzy Relative of the k-Medoids A lgorithm with Application to Document and Snippet Clustering, Proc. *IEEE Intl. Conf. Fuzzy Systems - FUZZ_IEEE 1999*., Korea, August 1999.
6. Relational Clustering Based on a New Robust Estimator with Applications to Web Mining, Proc. *Intl. Conf. North American Fuzzy Info. Proc. Society (NAFIPS 99)*, New York, June 1999.
7. Mining and Warehousing Web Logs, *Proc. Workshop on Web Information and Data Management, ACM Conf. on Information and Knowledge Management*Kansas City, MO, 1999.
8. On Mining Web Access Logs, Proc. SIGMOD 2000 Workshop on Research Issues in Data Mining and Knowledge Discovery, pp 63-69, Dallas, 2000.
9. Retriever: Improving Web Search Engine Results Using Clustering, to appear in *Managing Business with Electronic Commerce: Issues and Trends* (A. Gangopadhyay editor).
10. Low Complexity Fuzzy Relational Clustering Algorithms for Web Mining, accepted for publication in *IEEE Trans. Fuzzy Systems*, 2001

Area Background

The research under this project covers two areas -- Web personalization/mining, and robust fuzzy clustering. We briefly describe how we are applying techniques from these two areas in our project.

The idea of "personalizing" the Web space for users, in part by "mining" the Web documents and the users' traversals of the Web space, can trace its descent from the work in information retrieval and filtering. The problem can be attacked both from the client end by creating browsers that filter incoming information, and also with servers that adapt the information returned based on the user. The latter solution is further attractive since it also cuts down on network traffic. A related topic is the idea that we can learn much about users and customers by tracking and analyzing their *clickstreams,* which is of great importance in e-commerce. However, Web mining in general, and site analysis in particular, may well turn out to be a flop unless proper attention is paid to formalizing the goals, analysis, and evaluation methods.

Blindly dumping Web data into a data mining tool is not expected to yield good results. Any trends or patterns to be discovered, are buried in data with significant noise components. Hence, robust methods, that can deal with significant amounts of noise, are needed. Predictive models (including statistical tools) cannot produce reliable results unless the data represents a significant percentage of the total number of possible (traversal) patterns, which is astronomically large. Thus, methods for reducing the dimensionality of the problem are required. Also, many Web objects are non numeric in nature, making "distance" measurements and "equality" judgements an issue. Hence, algorithms that can also deal with non-numeric data are desirable. Moreover, the categories in the data are fuzzy. Our approach is to simplify the problem by categorizing the user space as well as the document space by applying low-complexity robust fuzzy clustering techniques.

Area References

This is a relatively nascent area. Besides the Web site we maintain for the project(http://www.cs.umbc.edu/~ajoshi/Web-mine/), other sites which reflect current work and provide good overviews are U. Washington (http://www.cs.washington.edu/research/adaptive/) , and Simon Fraser University (http://db.cs.sfu.ca/WebMiner/). Other groups working in this and related areas are at USC (Shahabi et al.) and Depaul University (Mobasher et al.)

Potential Related Projects

To the best of our knowledge, this is the only project in its area supported by IDM. There are some ongoing projects in data mining (P. Smith) and agents for Web browsing (S. Gauch) that are possibly relevent to this work.