A FORMAL CONCEPT ANALYSIS APPROACH FOR WEB USAGE MINING

Baoyao Zhou, Siu Cheung Hui and Kuiyu Chang
School of Computer Engineering, Nanyang Technological University, Singapore

Abstract: Formal Concept Analysis (FCA), which is based on ordered lattice theory, is applied to mine association rules from web logs. The discovered knowledge (association rules) can then be used for online applications such as web recommendation and personalization. Experiments showed that FCA generated 60% fewer rules than Apriori, and the rules are comparable in quality according to three objective measures.

Key words: web intelligence, knowledge discovery, data mining, knowledge-based systems, web usage mining, Formal Concept Analysis, association rules

1. INTRODUCTION

Web usage mining [1], also known as web log mining, aims to discover interesting and frequent user access patterns from web browsing data stored in the log files of web/proxy servers or browsers. The mined patterns can facilitate web recommendations, adaptive web sites, and personalized web search and surfing.

Various data mining techniques [2] such as statistical analysis, association rules, clustering, classification and sequential pattern mining have been used for mining web usage logs. Statistical techniques are the most prevalent; typical extracted statistics include the most frequently accessed pages, average page viewing time, and average navigational path length. Association rule mining can be used to find related pages that are most often accessed together in a single session. Clustering is commonly used to group users with similar browsing preferences or web pages with

semantically related content. Classification is similar to clustering, except that a new user (page) is classified into a pre-existing class/category of users (pages) based on profile (content). Sequential pattern mining involves identifying access sequences that are frequently exhibited across different users. All of the aforementioned techniques have been successfully deployed in various web-mining applications such as web recommendation systems [3], whereby web pages likely to be visited by users in the near future are recommended or pre-fetched.

Formal Concept Analysis (FCA) [4] is a data analysis technique based on ordered lattice theory. It defines a formal context in the form of a concept lattice, which is a conceptual hierarchical structure representing relationships and attributes in a particular domain. Formal concepts can then be generated and interpreted from the concept lattice using FCA. FCA has been applied to a wide range of domains including conceptual clustering [5], information retrieval [6], and knowledge discovery [7].

A novel web usage mining approach using FCA is proposed. In particular, association rules are extracted from web logs using FCA, which can efficiently and accurately identify frequent patterns.

2. FCA-BASED WEB USAGE MINING

Web usage data [2] can be extracted from the log files of web/proxy servers and browsers or any other data resulting from online interaction. Without loss of generality, only web server logs are considered here. Figure 1 gives an overview of the FCA-based web usage mining. The proposed approach consists of the following steps: (1) Preprocessing; (2) Web usage context construction; (3) Web usage lattice construction; and (4) Association rules mining.

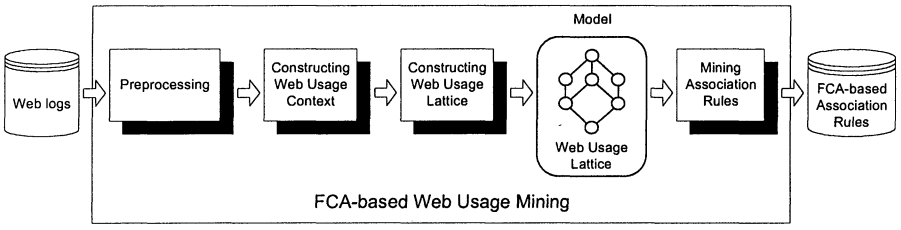

Figure 1. Overview of FCA-based web usage mining.

3. PERFORMANCE EVALUATION

The performance of FCA mined rules is benchmarked against the Apriori-based algorithm [8] on a web recommendation dataset.

3.1 Web Recommendation

Association rules are well suited for web recommendations. The recommendation engine generates recommendation (links) by matching users' recent browsing history against the discovered association rules. Three measures are defined to objectively evaluate the performance of FCA versus Apriori.

Definition Let N be the total number of rules, N_c the number of correct rules (true for the page accessed immediately after the current page), N_s the number of satisfactory (fired) recommendation rules (for pages accessed later in the same session), N_n the number of all nonempty recommendation rules, then we define three measures: $precision = \frac{N_c}{N}$ $satisfaction = \frac{N_s}{N}$ $applicability = \frac{N_n}{N}$.

Intuitively, precision measures how likely a user will access one of the recommended pages immediately. Satisfaction is a broader yet important measure that counts how likely a recommended page may be accessed in the future during the same session. That is because oftentimes the immediate following web page accessed by a user may be an interim navigational page instead of the target page. Applicability estimates how often recommendations will be generated.

3.2 Performance Results

Experiments were written in C++ and simulated on a 1.6 GHz machine. Two session datasets from Microsoft's Anonymous Web Data (http://kdd.ics.uci.edu) were used. This dataset records users' access in a one-week period during February 1998. To construct the web usage lattice, 2213 valid sessions out of 5000 were used, each with 2 to 35 page references (out of 294 pages). The test dataset contains 8,969 valid sessions (out of 32,711).

Figure 2(a) tallies the number of extracted association rules using a support of 10 and confidence values of 10% to 90% by AR and FCA. Clearly, FCA extracted far fewer rules (N_FCA) than Apriori (N_AR). Moreover, the precision, satisfaction, and applicability of FCA rules versus Apriori (AR) as shown in Figures 2 (b), (c), and (d) respectively are only marginally lower.

4. CONCLUSIONS

A Formal Concept Analysis (FCA) approach for web usage mining was proposed and evaluated against the classical Apriori algorithm. FCA generated 60% fewer rules than Apriori with comparable quality according to three objective measures. The FCA approach is thus an effective and efficient tool for generating web recommendations. Another benefit of FCA is that mined association rules can be visualized directly from the FCA lattice (not shown here for brevity).

REFERENCES

1. R. Kosala, and H. Blockeel, "Web Mining Research: A Survey", ACM SIGKDD Explorations, Vol. 2, 2000.
2. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM SIGKDD Explorations, Vol. 1. No. 2, 2000, pp. 12 - 23.
3. Ş. Gündüz, and M. T. Özsu, "Recommendation Models for User Accesses to Web Pages", In Proc. of 13th Intl. Conf. Artificial Neural Networks (ICANN), Istanbul, Turkey, June 2003, pp. 1003-1010.
4. R. Wille, "Restructuring lattice theory: an approach based on hierarchies of concepts", In I. Rival. Editor, Ordered sets. Boston-Dordrecht: Reidel, 1982, pp. 455-470.
5. G. Stumme, R. Taouil, Y. Bastide, and L. Lakhal, "Conceptual Clustering with Iceberg Concept Lattices", In: Proc. of GI-Fachgruppentreffen Maschinelles Lernen'01, Universität Dortmund, vol. 763, October 2001.
6. C. Lindig, "Concept-Based Component Retrieval", In Working Notes of the IJCAI-95 Workshop: Formal Approaches to the Reuse of Plans, Proofs, and Programs, August 1995, pp. 21-25.
7. J. Hereth, G. Stumme, U. Wille, and R. Wille, "Conceptual Knowledge Discovery and Data Analysis", In Proc. of ICCS2000, 2000, pp. 421-437.
8. R. Agrawal, and R. Srikant, "Fast Algorithms for Mining Association Rules" In Proc. of VLDB Conference, 1994, pp. 478-499.

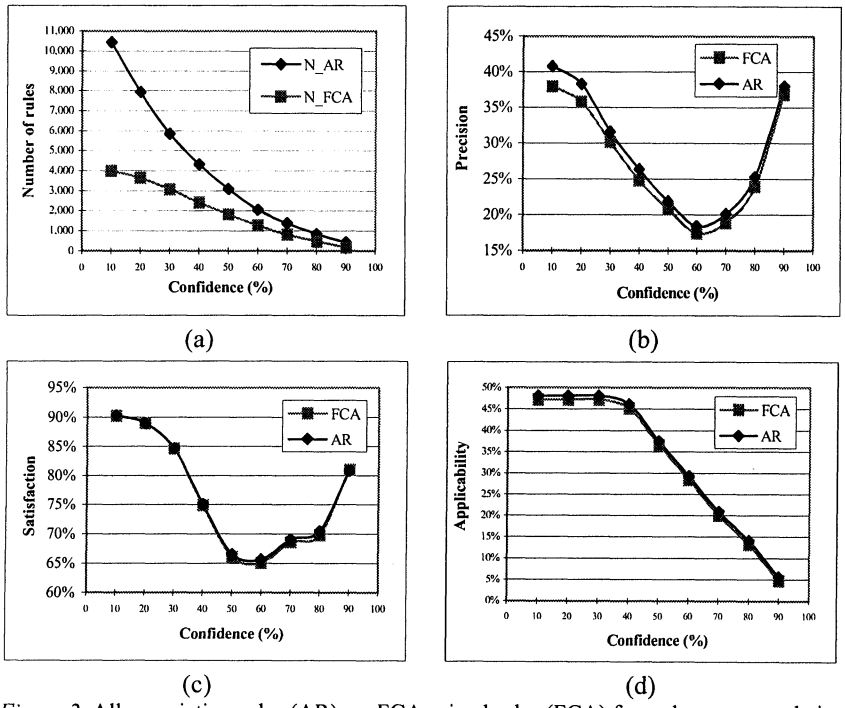

Figure 2. All association rules (AR) vs. FCA-mined rules (FCA) for web recommendation.