Association Rules Mining from Context-enriched Server Logs

Juraj VIŠŇOVSKÝ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
visnovsky.j@gmail.com

As users are browsing the Web, servers are recording millions of theirs actions to eventually offer better service. These large amounts of Web logs should be reused, otherwise their recording could be considered as a waste of resources. This field is covered by Web usage mining which discovers Web usage patterns. In our work we are going to prove the importance of context in the field of Web usage mining.

By using web usage mining techniques we are able to discover information about users' behaviour. Resolving their habits may be used in improving personalized recommendation and targeted advertising. For our purpose we are going to mine frequent patterns using FP-Growth algorithm and then generate association rules.

In this paper we focus on the analysis of Adaptive proxy server [2] access logs. The goal of the Adaptive proxy server is to improve user's experience by personalising Web content. Our dataset consists of more than 3 million access logs describing the activity of 77 unique users. When analysing server access logs, we have to bear in mind that these logs represent human actions and we have to consider many different contexts which could affect user's activity. Expressing what the word context means can be difficult and so is to define it. Probably the best definition came up from Dey [1]. According to this definition context is any information that can be used to describe a state of an entity. The entity could be a person, an object or a place that is relevant to the interaction between an application and a user.

We consider only small amount of contextual information influencing user's actions. As we are slightly limited by using the access logs, neither user's current mood nor his exact location is able to be found out. We use contexts as follows:

- Time
- Location and occupation
- Weather
- Web domain category

* Supervisor: Dušan Zeleník, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 51–52.

Our method of generating association rules analyses logs gathered by Adaptive proxy server. This data are being processed by our method in three steps as seen in Figure 1.

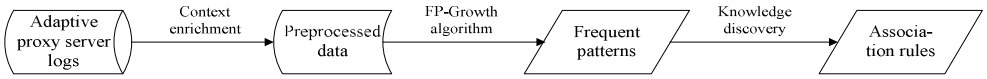

Figure 1. An overview of the method of association rules mining in access logs.

An association rule represents a correlation between two or more items which do not seem related at the first glance. An association rule can be expressed as logical implication A ⇒ B with attributes of support and confidence. Support is the probability both A and B occur and the probability that B occurs when A is already present is called confidence.

For discovering association rules we propose a modification in FP-Growth algorithm. While we are seeking to discover context-aware association rules, every node of the FP-Tree will be represented by a context-enriched access log.

Frequent pattern growth algorithm needs two scans of the access logs stored in database. At the first scan, algorithm evaluates occurrence of every context-enriched access log. Then the algorithm builds FP-Tree structure and inserts only the most frequent access logs as the tree nodes. Minimum support threshold defines how many times a log has to be noticed during the scan to be considered a frequent item.

FP-Growth algorithm handling large amounts of access logs may produce a high number of frequent patterns. The results may consist of strong related association rules. These can, however, be misleading. In order to improve our result-set we have to get rid of misleading, obvious or redundant association rules.

To evaluate our predictions of future events, the set of access logs will be split into two parts. The first larger part represents training interval of context-enriched access logs. Based on knowledge gathered from this data interval we are going to generate predictions of the future events. The prediction has to consist of the combination of context data and access logs, as we are not interested in uninteresting, misleading or obvious statements (e.g. “It is Saturday.” ⇒ “It is raining.”). The second part is a set of logs which will be used for calculating the accuracy of the prediction.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Dey, A. K.: Understanding and Using Context. *Personal and Ubiquitous Computing,* (2001), pp. 4-7.

[2] Kramár, T., Barla, M., Bieliková, M.: PeWeProxy: A Platform for Ubiquitous Personalization of the "Wild" Web. *Proceedings of the 19th International Conference on User Modelling, Adaptation and Personalization*, (2011), pp. 7-9.