Applying Web Usage Mining Techniques to Discover Potential Browsing Problems of Users

I-Hsien Ting, Chris Kimble and Daniel Kudenko
Department of Computer Science, The University of York
{derrick, kimble, kudenko}@cs.york.ac.uk

Abstract

In this paper, a web usage mining based approach is proposed to discover potential browsing problems. Two web usage mining techniques in the approach are introduced, including Automatic Pattern Discovery (APD) and Co-occurrence Pattern Mining with Distance Measurement (CPMDM). A combination method is also discussed to show how potential browsing problems can be identified

1. Introduction

Website design is an important criterion for the success of a website. In order to improve website design, it is essential to understanding how the website is used through analysing users' browsing behaviour. Currently there are many ways to do this, and analysis of the clickstream data is claimed to be the most convenient and cheapest method [3]. Web usage mining is a tool that applied Data Mining techniques to analyse web usage data [1], and it is a suitable technique that can be used to discover potential browsing problems. However, traditional web usage mining techniques are not sufficient enough for discovering potential browsing problems, such as Clustering, Classification and Association Rule.

In this paper, we proposed an approach, which is based on the concept of web usage mining and follows the *KDD (Knowledge Discovery in Database)* process. Two main techniques are included, which are Automatic Pattern Discovery, and a co-occurrence pattern mining, which is improved from traditional traversal pattern mining. These techniques are claimed can be used to discover potential browsing problems.

2. An Approach for Applying Web Usage Mining Techniques

In this paper, we proposed an approach for applying web usage mining techniques to discover potential browsing problems. Figure 1 presents the proposed approach, which is based on the KDD process [2]. In this approach, the KDD process will be run as a normal process, from data collection and pre-

processing, to pattern discovery and analysis, recommendation and action. The second step (pattern discovery and analysis) will be the main focus of this paper.
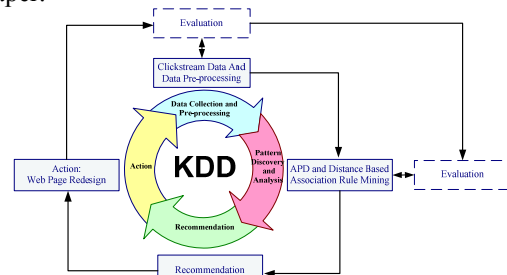

Figure 1. A KDD based approach for discovering potential browsing problems

3. Automatic Patterns Discovery (APD)

In our previous work [4], some interesting patterns have already been identified, including *Upstairs* and *Downstairs* pattern, *Mountain* pattern and *Fingers* pattern. The *Upstairs* pattern is found when the user moves forward in the website and never back to the web page visited before. The *Downstairs* pattern is that the user moves backward, that is the user returns to the visited pages. The *Mountain* pattern occurs when a Downstairs pattern immediately follows an Upstairs pattern. The *Fingers* pattern occurs when a user moves from one web page to browse another web page and then immediately returns to the first web page. These patterns are claimed to be very useful for discovering potential browsing problems (see [4] for further detail).

The APD method is based on the concept of sequential mining to parse the browsing routes of users. The APD method is performed by a three-level browsing route transformation algorithm. The level-1 elements include *Same*, *Up* and *Down*. The level-2 elements are *Peak* and *Trough*, and the final level is to discover the *Stairs*, *Fingers* and *Mountain* pattern (See [5] for more detail about the APD method).

Table 1 shows an example of number-based browsing sequences, which are transformed from the browsing routes of users (the number denotes the

occurrence sequence of the visited web page in a user's session). Table 2 shows the discovered final patterns by performing the APD method.

Table 1. Number-based browsing sequences

Number	Number-based sequence
1	0,1,2
2	0,0,1,0,2,0,3,0,4,0,5,6,7,6,7,8,6,4,6,5,0

Table 2. Final patterns

Number	Patterns
1	Upstairs
2	Finger, Finger, Finger, Finger, Mountain, Mountain, Mountain

4. Co-occurrence Pattern Mining with Distance Measurement (CPMDM)

CPMDM is another technique that can be used to analyse the browsing behaviour of users, which is an improvement of co-occurrence pattern mining by introducing a *Distance* measurement. Co-occurrence pattern is a pattern that used to describe the co-occurrence frequency (or probability) of two web pages in users browsing routes. The additional measurement, **Distance**, is a measurement that used to measure how many browsing steps from one page to another in a co-occurrence pattern.

There are three different directions of the distance measurement, including **Forward**, **Backward** and **Two-Way**. The **Forward** distance measures the distance from web page A to B of the co-occurrence pattern A→B. The **Backward** distance on the other hand measures the distance from B to A of the co-occurrence pattern A→B. The **Two-Way** distance combines forward and backward distance. It ignores the direction of the association rule, and takes all co-occurrence patterns about A and B.

5. Combining APD and CPMDM for Discovering Browsing Problems

The analysis results of the APD and CPMDM are two totally different analyses of users' browsing behaviour. However, there will be some biases if only one of these two methods is used to assess the website's design. Therefore, if the analysis results of the APD and CPMDM can be combined, more concrete indications of potential problems in the website's design can be discovered.

Table 3 shows an example about combining the APD and CPMDM methods for discovering potential browsing problems. In this case, the starting page of the co-occurrence patterns is the home page of the

University of York website. In the table, the Support means the probability of the co-occurrence pattern and the Distance is the average forward distance of the pattern. The proportion of Stairs and Fingers pattern is measured by using the APD method. In this case, we consider that the fingers pattern is a problematical pattern, and the longer the distance means the more difficult for a user to traverse from one page to another. Therefore, the browsing route from home page to /uao/ugrad/course page can easily to be discovered as a route that potential browsing problem may occur.

Table 3. Combining the APD and CPMDM of "The people who view home page then view"

URL	Support	Distance (Average)	Stairs Pattern	Fingers Pattern
/uao/ugrad/	0.25	9.1271	44%	39%
/gso/gsp/	0.173	5.3195	52%	26%
/uao/ugrad/courses/	0.127	**16.9021**	34%	**47%**

6. Conclusion

This paper proposed a users' browsing behaviour analysis approach which is based on applying web usage mining techniques. The concepts of the APD and CPMDM have been briefly introduced, and the combination method has been discussed in this paper as well. From the example of the combination method, it showed that potential browsing problems of users can be discovered easily. The approach that proposed in this paper is therefore beneficial for the area of website design improvement.

7. References

[1] Cooley, R. Mobasher, B. and Srivastave, J. (1997) "Web Mining: Information and Pattern Discovery on the World Wide Web" *In Proceedings of the 9th IEEE ICTAI Conference*, pp. 558-567, Newport Beach, CA, USA

[2] Lee, J., Podlaseck, M., Schonberg, E., Hoch, R. (2001) "Visualisation and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising", *Journal of Data Mining and Knowledge Discovery*, Vol. 5, pp. 59-84

[3] Kohavi, R., Mason, L. and Zheng, Z. (2004) "Lessons and Challenges from Mining Retail E-commerce Data" *Machine Learning*, Vol. 57, pp. 83-113

[4] Ting, I. H., Kimble, C., Kudenko, D. (2004) "Visualizing and Classifying the Pattern of User's Browsing Behavior for Website Design Recommendation" *In Proceedings of 1st KDDS workshop,* 20-24 September, Pisa, Italy

[5] Ting, I., Clark, L., Kimble, C., Kudenko, D. and Wright, P. (2007) "APD-A Tool for Identifying Behavioural Patterns Automatically from Clickstream Data" *Accepted to appear in KES2007 Conference*, 12-14 September