An Efficient Multidimensional Data Model for Web Usage Mining

Edmond H. Wu[1], Michael K. Ng[1], and Joshua Z. Huang[2]

[1] Department of Mathematics, The University of Hong Kong
hcwu@hkusua.hku.hk,mng@maths.hku.hk
[2] E-Business Technology Institute,The University of Hong Kong
jhuang@eti.hku.hk

Abstract. Web applications such as personalization and recommendation have raised the concerns of people because they are crucial to improve customer services, particularly for E-commerce Websites. Understanding customer preferences and requirements in time is a premise to optimize these Web services. In this paper, a new data model for Web data is introduced to analyze user behavior. The merit of the cube model is that it not only aggregates user access information but also takes the Web structure information into account. Based on the model, we propose some solutions to intelligently discover interesting user access patterns for Website optimization, Web personalization and recommendation. We used the Web usage data from a sports Website in China to evaluate the effectiveness of the model. The results show that this integrated data model is effective and efficient to apply into practical Web applications.

1 Introduction

Understanding the user behavior is the first step to provide better Web services. Under this premise, technologies for Web applications, such as user profiling, personalization and recommendation systems are frequently used. Hence, developing effective and efficient solutions to discover user patterns and then to optimize Web services has become an active research area in Web usage mining.

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications [4]. In [4], J. Srivastva et. al also propose a three-step Web usage mining process which are called preprocessing, pattern discovery, and pattern analysis.

One of the new trends in Web usage mining is to develop Web usage mining system that can effectively discover users' access patterns and then intelligently optimize the Web services. Recent studies[1, 3, 5]have suggested that the structural characteristics of Websites, such as the Website topology, have a great impact on the performance or efficiency of Websites. Hence, combining with the structure of Website, we can gain more interesting results for Web usage analysis.

In this paper, we present an efficient data model for aggregating user access sessions to effectively support different mining tasks. Based on the data

J.X. Yu, X. Lin, H. Lu, and Y. Zhang (Eds.): APWeb 2004, LNCS 3007, pp. 373–383, 2004.

model, we can easily perform different mining tasks, such as association rule mining, sequential pattern mining, clustering, and Web usage predicting etc. Using these mining results, we can provide multiple solutions for various Web applications. For example, personalized services, effective recommendation system, Web prefetching and Website optimization.

This paper is organized as follows. In Section 2, we present the framework of a multidimensional Web usage mining model. In Section 3, we introduce the implementation of the model, after that, experiment results are given. Then, we demonstrate practical Web applications in a real Website for optimizing Website services based on the model in Section 4. Finally, We give some conclusions and present our future work in Section 5.

2 A Multidimensional Model for Web Usage Analysis

From above introduction, we can see that Web usage mining techniques have been widely used in various Web applications. Hence, it is necessary to develop an integrated platform for effective Web usage analysis. For this reason, we propose a multidimensional model to smoothly integrates Web data preprocessing, Website content and topology information. The framework can also easily combine different data mining algorithms to support different Web applications. In this section, we will introduce the main components of the model individually.

2.1 Preprocessing Module

In [7], we introduced a data-cube model to contain the original access sessions for data mining from Web-logs. Based on it, we also investigated the practice of dealing with Web-log data streams[6]. These work provided feasible preprocessing solutions to turn large volumns of Web logs into useful session information. So, the model designed can support both online and offline Web usage analysis.

2.2 The PUT-Cube Module

The PUT-Data Cube is the core of the multidimensional model. As we have mentioned, the Website structure information can greatly help us to analyze user patterns. Hence, we propose a data cube model which integrates Website topology, content, user and session information for multiple usage mining tasks. A novel cube model called PUT-Cube is defined as follows:

Definition 1. A PUT-Cube model is a four tuple $\langle P, U, T, \mathcal{V} \rangle$ where P, U, T are the sets of indices for three main dimensions(Page, User, Time) in which

1. P indexes all page related attributes $P = P_1, P_2, ..., P_n$.
2. U represents of all user attributes $U = U_1, U_2, ..., U_m$ identifying users of groups or individual.
3. T indicates a set of temporal related attributes, $T = T_1, T_2, ..., T_r$, each describing the occurrence time or duration of user accesses.
4. $\mathcal{V}$ is a bag of values of all attribute combinations.

The PUT-Cube model focuses on most important factors in Web usage mining. If also considering the Website topology, the page factor not only provides information about which pages have been accessed, but also provides the information about their access order and relative positions. The user and time factors suggests who and when involve the Web access events. Therefore, based on the 'who' ,'when' and 'which' user access information from the PUT-Cube, we can discover more useful user patterns, and then try to explain 'why'. This mining process is desired because we have concentrated on the primary factors for analyzing user behavior. The PUT-Cube model also provides the flexibility of selecting relevant dimensions(or attributes) for analysis.

2.3 Algorithm and Application Module

There are different kinds of data mining algorithms which can adapt in Web usage mining, such as association rule mining and clustering. For example, in [5], we proposed a graph-based algorithm to find interesting association rules based on the Website topology. Our multidimensional model is capable of integrating these data mining algorithms efficiently. We can also set the thresholds in the data mining models to see different mining results. Moreover, we can compare the results from different data mining algorithms for validation and deeper analysis. Then, we can improve the Website services by purposing different Web applications. As results, the Website content and topology should be changed to meet the needs of visitors. The effectiveness of the changes can also be validated under this framework. Fig 1. shows the workflow of the multidimensional model. We can see that the PUT-Cube module plays a key role in organizing the Web usage analysis.

3 Implementation and Experiments

3.1 Model Implementation

In order to adapt the needs of different applications, we can set concept levels or specify value range to tailor a suitable model. For example, we can classify the Web pages into different categories, such as index page or content page. We may not need to include all the visitors in the 'U' factor. In fact, we usually select those users related to the user behavior analysis tasks, such as users who visit frequently. Also, we prefer to set different temporal levels in the occurrence time T_I dimension, such as minute, hour, day and month, which also depend on the analysis tasks.

The following is an example of the model implementation. First, we select dimensions related to P-U-T factors. For example, in order to get the access matrix(like Table 2), we select P_1, P_2 where $P_1 = P_2 = V$, V is the set of Web pages in the Website topology. Each entry $C(i, j)$ represents the number of accesses from V_i to V_j. As to 'U', we adopt the set of User IDs for analysis. As to 'T', we set two different temporal dimensions. One suggests the time spent

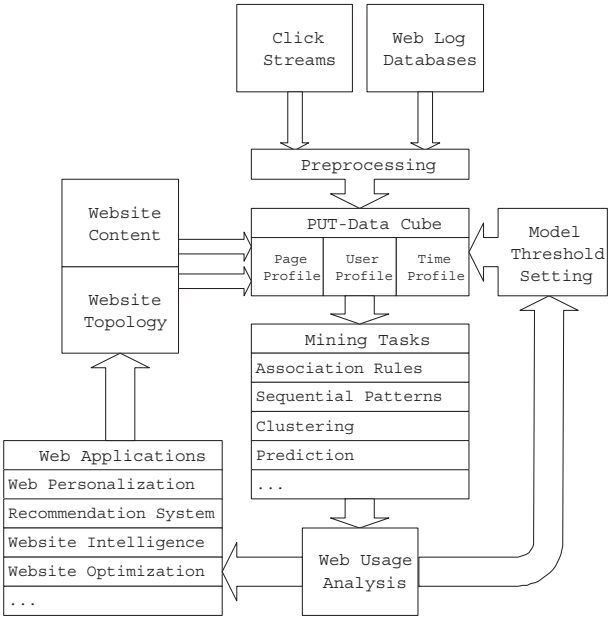

Fig. 1. The Multidimensional Model for Web Usage Mining

at the pages concerned. Another records an access occurrence time during given time intervals(such as minute, hour etc).

Based on above analysis, we can implement a PUT-Cube model. The first is to aggregate the sessions into a multidimensional data cube model. Given a access session $S = S_1, S_2, ..., S_n$ by User k during time interval l, $T = T_1, T_2, ..., T_n$ index the time spent at each page in the session. The current Website topology is $G = \{V, E\}$, the corresponding topology probability matrix is P(refer to [5]). The session will be aggregated into the cube by the following operations:
(1)As to page access matrix, For each access S_iS_j,$C_k^l(S_i, S_j) = C_k^l(S_i, S_j) + 1$;
(2)As to page probability matrix, $P_k^l(S_i, S_j) = P(S_j|S_i)$;
(3)As to time duration matrix, $T_k^l(S_i, S_j) = T_i + T_j$;

$C_{ij} = Count(S_i, S_j)$ indicates the number of accesses from S_i to S_j. We notice that the access session is the traversal path that a user follows in a Website topology. Any two adjacent pages in the session must be accessible by one click. Therefore, there are two cases of such counting, one is that S_i and S_j are adjacent in the access session S which also means the link of $SiSj \in E$. The other case is that S_i and S_j are adjacent which means there exist at least one page between S_i and S_j. For the first case, we will count it whenever it appears. As to the second case, we will only count once in the access session. Here is an example.

Example 1: Given access session $S = \{A, B, C, B\}$, in the first case, we will count AB, BC, and CB once since they are adjacent in the session, respectively.

As to the second case, we count AC once. But we won't count AB again since it has been counted once in the first case. So, $C(A,B) = 1$, $C(A,C) = 1$, $C(B,C) = 1$, $C(C,B) = 1$.

3.2 Cube Operations

Based on the cube model, we can explore our mining tasks to analyze the user behavior. we can select related dimensional attributes to form a cuboid lattice to discover interesting patterns. Cube operations, such as roll-up, drill-down, slice and dice can be easily performed in the data cube. Hence, the cube model is quite efficient to support different OLAP queries. Here is an example.

Example 2: Table 1 is a sample user access session dataset which contains the access sessions of three users in six different time intervals in given site topology(See the left topology of Fig 2). Table 2 shows the aggregating matrix which sums up the accesses of three users during the specified time intervals. From the matrix, we observed that the most frequent access is AC which has total 9 accesses.

Table 1. User Access Sessions

	User1	User2	User3
Time1	EHC	ABFBG	ABGB
Time2	ABF	ACHEC	ACH
Time3	ACHE	GBA	ACHEHC
Time4	AD	CH	ABF
Time5	DACH	ABFBG	ABGB
Time6	EH	ADACHEHC	ACHC

Table 2. Aggregating Sessions in Matrix

Page	A	B	C	D	E	F	G	H
A	0	7	9	2	5	4	4	8
B	3	0	3	0	0	4	5	3
C	4	4	0	0	5	4	2	8
D	3	1	3	0	3	1	1	3
E	4	5	4	0	0	4	2	4
F	0	2	0	0	0	0	2	0
G	3	3	3	0	0	0	0	3
H	4	4	6	0	5	4	1	0

3.3 Model Performance Analysis

We used the Web usage data from ESPNSTAR.com, a sports Website in China, to test and evaluate the performance and effectiveness of our PUT-Cube Model proposed. All the experiments were performed on a PC with a Pentium4 2.0GHz CPU and 256M main memory, running on Microsoft Windows 2000 Professional.

With permission, we got the topology of the Website for analysis. We use two months Web log data to do the experiments. The original Web logs contain millions of access records from the Web servers. After data preprocessing, we got the user access session datasets. In some datasets, we take all the sessions during a period of time (e.g., one day). We also select the sessions from particular users for analysis in some datasets. Table 3 shows some of the datasets for experiments. ES1, ES2 and ES3 are the access session datasets from the logs during December, 2002 and ES4 and ES5 are the logs from April, 2003.

We first evaluate the efficiency of aggregating access sessions by PUT-Cube. In this experiment, we use dataset ES3 to test the execution time when increasing

Table 3. Real Datasets

Dataset	No.Accesses	No.Sessions	No.Visitors	No.Pages
ES1	583,386	54,300	2,000	790
ES2	2,534,282	198,230	42,473	1,320
ES3	6,260,840	517,360	50,374	1,450
ES4	78,236	5,000	120	236
ES5	7,691,105	669,110	51,158	1,609

the input access sessions. From the result (See Table 4), we can see that the increase of running time exhibits a linear relationship with the increase of input access sessions.

Another experiment is to evaluate the efficiency of the PUT-Cube when increasing the cardinality. We use datasets ES 1, 4, and 5 for testing. The result shows in Table 5. Through above analysis, we noticed that even for a Website with thousands of different pages, users or time intervals, the PUT-Cube model is still feasible. We can get the results in an acceptable period of time.

The third experiment is to test the efficiency of the PUT-Cube model when increasing the number of dimensions. We choose the largest dataset ES5. The result shows in Table 6. The result shows that the cube model still work well even we include most of the session attributes in the cube model. In practice, we usually choose two to five dimensions for analysis. Hence, the cube model can work well for most analysis tasks, even for online analysis.

Table 4. Running Time Vs # of Sessions

Number	10000	50000	200000	400000
Runtime(S)	11	45	206	480

Table 5. Running Time Vs # of Cardinality

Number	500	1500	3000	5000
Runtime(S)	25	120	293	605

Table 6. Running Time Vs # of Dimensions

Number	3	5	7	9
Runtime(S)	15	71	320	945

4 Web Applications

In this section, we demonstrate several practical application based on the multidimensional model combining different data mining algorithms proposed.

4.1 Website Optimization

We first propose our solution for Website optimization. In this paper, Website optimization refers to reorganize the Website topology and content to improve the access efficiency and system performance.

Measure of Discovery of Access Patterns

In [6], we suggest a novel measure named Access Interest (AI) as below:

Definition 2. Given Aggregating Session Matrix C and Topology Probability Matrix P, page staying time T, the Access Interest Matrix is given by:

$$AI(i,j) = In(\frac{C_{ij}T_{ij}}{P_{ij}} + 1) \quad (1)$$

where C_{ij} is the number of accessing from Web page V_i to V_j, and P_{ij} is the probability from V_i to V_j, T_{ij} is the average stay time of visiting V_i and V_j. Given $T = T_1, T_2, ..., T_n$, T_i is the average staying time of V_i. We define $T_{ij} = T_i + T_j$.

Example 3: Using the sample dataset in Table 1, we discovered top 5 AI values in the given site topology(See the left topolgoy in Fig 2) . They are: $AI(E,F) = 13.1, AI(H,F) = 12.5, AI(G,H) = 11.7, AI(E,G) = 11.4, AI(C,F) = 11.2$, respectively. From the results, we can conclude that these access patterns are important, however, the access efficiency is not satisfying. Discovering such patterns will be of great help to improve the Website access efficiency no matter from the user or system point of views. Hence, the results validated the effectiveness of using the cube model to mine exceptional user access patterns.

Algorithm for Website Optimization

The Website topology optimization procedure contains four major steps. We summarize them as follows: The first step is to mine the interesting access pattern model described above. In the second step, using the interestingness measures AI to identify interesting access patterns for analysis. In third step, optimizing the Website topology based on the interesting access patterns discovered. In the last step, we can choose UAE and SAE to validate the effectiveness of the optimization results(refer to [6]).

Since we have found some interesting access patterns which represent the low access efficiency, we can build some direct hyperlinks to connect them. For instance, if we found E to F is a low access efficiency pattern and there is no direct hyperlink between them, then we can build a new link in page E to F. As results, the access efficiency from E to F will be greatly improved. Here is the description of the algorithm:

Begin

1. Input Web access sessions database D and Website topology $G = (V, E)$
2. Mining interesting access patterns and gain AI matrix using PUT-Cube model
3. Input top k access patterns $P_1Q_1, P_kQ_k$with highest AI values and top m access patterns $P_1R_1, P_kR_m$with lowest access frequency $C(P_j, R_j)$
4. For each pattern $PiQi$, $i = 1, ..., k$, if $P_iQ_i \notin E$, build hyperlink P_iQ_i to connect the sequential access pattern. For each pattern $PjRj$, $j = 1, ..., m$, if $P_jR_j \in E$ and $AI(P_j, R_j)$ is low , disconnect hyperlink P_jR_j.
5. If the access efficiency of user patterns is satisfying, output the optimized Website topology

End

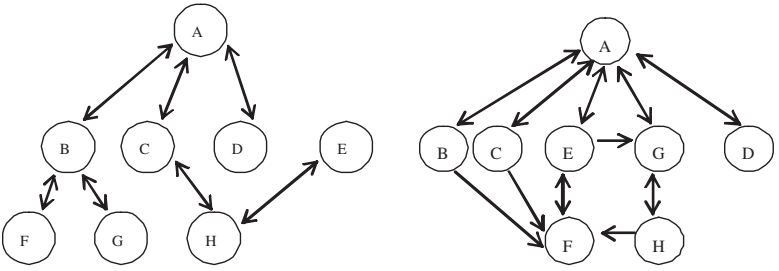

Fig. 2. Comparison of Original and Optimized Website Topology

Example 4: Fig. 2 shows the optimized topology returned by the topology optimization algorithm with linkage optimization. Comparing with the original topology (See the left of Fig 2), there are several significant changes in the structure of the topology. Recall the top 5 interesting access patterns EF, HF, GH, EG and CF, which can not be efficiently accessed in the original topology, however, in the new topology, they are quite efficient for the users to access.

In the optimized topology(See the right of Fig 2), B and C have the common linkage, if we merge B and C into a new page, the number of the links required in the optimal topology is approximately equivalent to the original one.

We should note that the optimization algorithm is feasible to apply in both global optimization and local optimization. That is, we can optimize the whole Website or part of the topology concerned.

A Real Case Study

We investigated the effectiveness of Website optimization model through a real case study. The server logs and Website topology is from ESPNSTAR.com.cn, a well-known E-commerce sports Website in China. The professional sports website has a focused set of interesting topics related to sports, such as football, basketball, and even golf. Hence, it has a very broad audience. An interesting discovery is that many visitors tend to visit the Website during some important matches, such as English Premier League, to look for related content. However, their access patterns change very quickly. As results, the problem occured was that large volumes of page requested made the Web serve unstable, and hence lower the access speed. Through our analysis about Website optimization, we note that unnecessary pages requested can be decreased if the Website topology can well adapt the current user access patterns. By the using Website topology algorithm based on the cube model, it is supposed that we can improve the access efficiency of Website.

Fig 3 is the comparison of access efficiency measures by original and optimized Website topology. We employed the ES5 dataset to do the experiment. The time period is from 0:00 AM to 4:00 AM, April 1, 2003. We can see that both the access efficiency measures of the optimized Website topology have great improvement than the original one, especially in the peak hours around 2 AM to 3 AM. It can be explained that during the peak hours, more people follow similar

access patterns, for example, searching for hot news, on-going matchs etc. The optimization model can help the Website to achieve more stable access efficiency. On average, it can have over 50% of improvement in access efficiency. That is, people may save a lot of time to search the content that they are interested in from the Website. The experiment validated the effectiveness of Website optimization based on our multidimensional model for Web usage mining. Hence, the model proposed is feasible to put into practice for Website optimization.

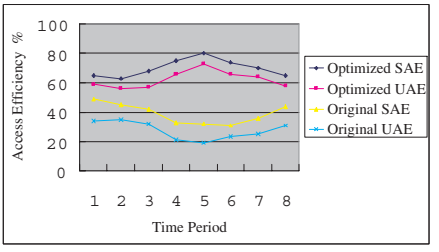

Fig. 3. Connectivity Optimization

Fig. 4. Personalized Web Services

4.2 Website Personalization

Other Web applications, such as personalization can also employ the multidimensional model to optimize the services. For example, we can use the cube model to find the access habits of a particular group of users or individual. With the help of domain experts, we purposed different cuboids to investigate user behavior. Several analytical cuboids are listed as follows:

- Cuboid_1:{Page_Category,Time_Occurrence, User_ID}
- Cuboid_2:{Page_Frequency,Time_Interval, User_Category}
- Cuboid_3:{Page_Frequency,Page_Topology, Time_Duration,User_Category}
- Cuboid_4:{Page_Frequency,Time_Occurrence, Time_Duration,User_ID}

These cuboids can analyze the Website access situations from different point of views. In our investigation to a sports Website, we found that as to Website analysts, they were quite eager to know which content pages are frequently visited during which period of time, particular user prefers to what kind of sports content. The cube model can achieve these analysis tasks easily. As to Cuboid_1, we can analyze a particular user's visiting preference varying page contents and time occurrence. Cuboid_2 provides the basic user access statistics of the Website. Employing Cuboid_3, we can monitor the user patterns and access efficiency of Website. Using Cuboid_4, we can employ other data mining techniques, such as cluster analysis and temporal association rules.

Based on the mining results, the personalization system based on the model can intelligently build a personalized Website based on the user's interesting access patterns. For example, the personalization system can intelligently post the headlines which the user will probably be interested in. For football fans, the personalized system can automatically select the most exciting football news for them in the headlines(See the pointer 1 of Fig 4).

4.3 Recommendation System

The model can also be employed in recommendation system. For example, we found an interesting access pattern that many visitors in the Website would like to browse the scoreboards and calendar of European football leagues very frequently, e.g., English Premier League. However, the related contents scatter in the original Website. Discovering such pattern, the Website designers put related content in the homepage, so users can easily to acquire these information(See the pointer 2 of Fig 4). Based on the individual access patterns, we can also promote some new sports activities or sports products to the users who may be interested in (See the pointer 3 of Fig 4).

Some suggestions have been adopted in the Website redesign(Fig 4 suggests the redesign of homepage which provides Web personalization and recommendation services). The dataset ES 4 contains the access activities of 120 members to their fee-paid services. By analyzing the customer behavior based on the cube model, the Website managers and designers can reorganize the content and its structure to provide better services. The cube model can also be employed in marketing research, CRM analysis etc. As results, these Web services can greatly improve customers' satisfaction.

5 Conclusion

In this paper, we propose an efficient multidimensional data model for combining user access sessions with Website topology for Web usage analysis. The model focuses on the page, user and time attributes to form a multi-dimension cube which can be frequently updated and queried. The experiments show that the data model is effective and flexible for different analysis tasks. Our real case studies suggest that the Web usage mining model can be applied in different Web applications, such as Website optimization, Web personalization and recommendation systems. In the future, we intend to implement the multidimensional model with more data mining algorithms and Web applications in a business intelligence platform.

Acknowledgements. We would like to thank Dr. Chris Ding, Computer Scientist of Lawrence Berkeley National Laboratory, for fruitful discussions on this topic. Thanks to Dr. Z. Lin, CTO of ESPNSTAR Multimedia of China, for his support to our work. This research was supported by the RGC Grant Nos. 7046/03P and HKU 7130/02P.

References

1. Bettina Berendt, Bamshad Mobasher, Miki Nakagawa, and Myra Spiliopoulou, The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis, *Proceeding of the WEBKDD 2002 Workshop*, Edmonton, Canada, 2002.
2. Robert Cooley, Pang-Ning Tan, and Jaideep Srivastava. Websift: The web site information filter system.*In Proceedings of the Web Usage Analysis and User Profiling Workshop*, 1999.
3. Miki Nakagawa, Bamshad Mobasher, A Hybrid Web Personalization Model Based on Site Connectivity, *WEBKDD*, 2003.
4. J. Srivastava, R. Cooley, M. Deshpande, and P. N. Tan. Web Usage Mining: Discovery and applications of usage patterns from web data.*SIGKDD Explorations,* 1:12-23, 2000.
5. Edmond H. Wu, Michael, K. Ng, A Graph-based Optimization Algorithm for Website Topology Using Interesting Association Rules, *Proc. the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2003)* ,Seoul, Korea, 2003.
6. Edmond H. Wu, Michael, K. Ng, and Joshua Z. Huang, On improving website connectivity by using web-log data streams, *Proc. of the 9th International Conference on Database Systems for Advanced Applications (DASFAA 2004)*, Jeju, Korea, 2004.
7. Q. Yang, J. Huang and M. Ng, A data cube model for prediction-based Web prefetching, *Journal of Intelligent Information Systems*, 20:11-30, 2003.