Web Usage Mining Consideration of General and Customised Access Patterns

R. Suganya[1]

[1]Department of Computer Science

[1]Bharathidasan University, Trichy, India

Abstract— With the continued growth of e-commerce, Web services, and Web-based information systems, the volumes of clickstream and user data collected by Web-based organizations in their daily operations has reached astronomical proportions. Analysing such data can help these organizations determine the life-time value of clients, design cross-marketing strategies across products and services, evaluate the effectiveness of pro-motional campaigns, optimize the functionality of Web-based applications, provide more personalized content to visitors, and find the most effective logical structure for their Web space. This type of analysis involves the automatic discovery of meaningful patterns and relationships from a large collection of primarily semi-structured data, often stored in Web and applications server access logs, as well as in related operational data sources

Key words: e-commerce, Web services, Web-based information systems, semi-structured data

I. Introduction

Web usage mining refers to the automatic discovery and analysis of patterns in clickstream and associated data collected or generated as a result of user interactions with Web resources on one or more Web sites [1,2]. The goal is to capture, model, and analyse the behavioural patterns and profiles of users interacting with a Web site. The discovered patterns are usually represented as collections of pages, objects, or re-sources that are frequently accessed by groups of users with common needs or interests.

II. Process Of Web Usage Mining

Web usage mining process can be divided into three interdependent stages: data collection and pre-processing, pattern discovery, and pattern analysis. In the pre-processing stage, the clickstream data is cleaned and partitioned into a set of user transactions representing the activities of each user during different visits to the site[2]. Other sources of knowledge such as the site con-tent or structure, as well as semantic domain knowledge from site ontolo-gies, may also be used in pre-processing or to enhance user transaction data. In the pattern discovery stage, statistical, database, and machine learning operations are per-formed to obtain hidden patterns reflecting the typical behaviour of users, as well as summary statistics on Web resources, sessions, and users. In the final stage of the process, the discovered patterns and statistics are further processed, filtered, possibly resulting in aggregate user models that can be used as input to applications such as recommendation engines, visualization tools, and Web analytics and report generation tools.

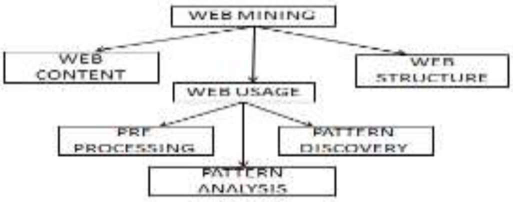


Fig.1: Classification of web mining

III. Data Preprocessing

An important task in any data mining application is the creation of a suit-able target data set to which data mining and statistical algorithms can be applied. This is particularly important in Web usage mining due to the characteristics of clickstream data and its relationship to other related data collected from multiple sources and across multiple channels. The data preparation process is often the most time consuming and computationally intensive step in the Web usage mining process, and often requires the use of special algorithms and heuristics not commonly employed in other do-mains. This process is critical to the successful extraction of useful pat-terns from the data. The process may involve pre-processing the original data, integrating data from multiple sources, and transforming the integrated data into a form suitable for input into specific data mining operations. Collectively, we refer to this process as data preparation[3].

Much of the research and practice in usage data preparation has been focused on pre-processing and integrating these data sources for different analysis. Usage data preparation presents a number of unique challenges which have led to a variety of algorithms and heuristic techniques for pre-processing tasks such as data fusion and cleaning, user and session identification, pageview identification [4]. The successful application of data mining techniques to Web usage data is highly dependent on the correct application of the pre-processing tasks. Furthermore, in the context of e-commerce data analysis, these techniques have been extended to allow for the discovery of important and insightful user and site metrics [5].

It provides a summary of the primary tasks and elements in usage data pre-processing. We begin by providing a summary of data types commonly used in Web usage mining and then provide a brief discussion of some of the primary data preparation tasks.

IV. Pattern Discovery

Usage data pre-processing results in a set of n pageviews, $P = \{p_1, p_2, \cdots, p_n\}$, and a set of m user transactions, $T = \{t_1,t_2,\cdots,t_m\}$, where each t_i in T is a subset of P. Pageviews are semantically meaningful entities to which mining tasks are applied (such as pages or products). Conceptually, we view each transaction *t* as an *l*-length sequence of ordered pairs:

$$t = \left\langle (p_1^t, w(p_1^t)), (p_2^t, w(p_2^t)), \cdots, (p_l^t, w(p_l^t)) \right\rangle,$$

where each $p_i^t = p_j$ for some *j* in {1, 2, ⋯, *n*}, and $w(p_i^t)$ is the weight associated with pageview p_i^t in transaction *t*, representing its significance. The weights can be determined in a number of ways, in part based on the type of analysis or the intended personalization framework. For example, in collaborative filtering applications which rely on the profiles of similar users to make recommendations to the current user, weights may be based on user ratings of items. In most Web usage mining tasks the weights are either binary, representing the existence or non-existence of a pageview in the transaction; or they can be a function of the duration of the pageview in the user's session. In the case of time durations, it should be noted that usually the time spent by a user on the last pageview in the session is not available. One commonly used option is to set the weight for the last page-view to be the mean time duration for the page taken across all sessions in which the pageview does not occur as the last one. In practice, it is common to use a normalized value of page duration instead of raw time duration in order to account for user variances. In some applications, the log of pageview duration is used as the weight to reduce the noise in the data[6].

For many data mining tasks, such as clustering and association rule mining, where the ordering of pageviews in a transaction is not relevant, we can represent each user transaction as a vector over the *n*-dimensional space of pageviews. Given the transaction *t* above, the transaction vector t (we use a bold face lower case letter to represent a vector) is given by:

$$\mathbf{t} = \left(w_{p_1}^t, w_{p_2}^t, \cdots, w_{p_n}^t\right),$$

where each $w_{pi}^t = w(p_j^t)$, for some *j* in {1, 2, ⋯, *n*}, if p_j appears in the trans-action *t*, and $w_{pi}^t = 0$ otherwise. Thus, conceptually, the set of all user trans-actions can be viewed as an *m*×*n*user-pageview matrix (also called the transaction matrix), denoted by *UPM.*

Given a set of transactions in the userpageview matrix as described above, a variety of unsupervised learning techniques can be applied to obtain patterns. These techniques such as clustering of transactions (or sessions) can lead to the discovery of important user or visitor segments. Other techniques such as item clustering and association or sequential pattern mining can find important relationships among items based on the navigational patterns of users in the site.

$$\mathbf{p} = \left(fw^p(f_1), fw^p(f_2), \ldots, fw^p(f_r)\right)$$

Where$fw^p(f_j)$ is the weight of the *j*th feature (i.e., f_j) in pageview *p*, for $1 \le j \le r$. For the whole collection of pageviews in the site, we then have an *n*×*r*pageview-feature matrix *PFM* = {p_1, p_2, …,p_n}. The integration process may, for example, involve the transformation of user transactions (in user-pageview matrix) into "content-enhanced" transactions containing the semantic features of the pageviews. The goal of such a transformation is to represent each user session (or more generally, each user profile) as a vector of semantic features (i.e., textual features or concept labels) rather than as a vector over pageviews. In this way, a user's session reflects not only the pages visited, but also the significance of various concepts or con-text features that are relevant to the user's interaction.

While, in practice, there are several ways to accomplish this transformation, the most direct approach involves mapping each pageview in a trans-action to one or more content features. The range of this mapping can be the full feature space, or feature sets (composite features) which in turn may represent concepts and concept categories. Conceptually, the transformation can be viewed as the multiplication of the user-pageview matrix *UPM*, de-fined earlier, with the pageview-feature matrix *PFM.* The result is a new matrix, *TFM* = {t_1, t_2, …,t_m}, where each t_i is a *r*-dimensional vector over the feature space. Thus, a user transaction can be represented as a content feature vector, reflecting that user's interests in particular concepts or topics.

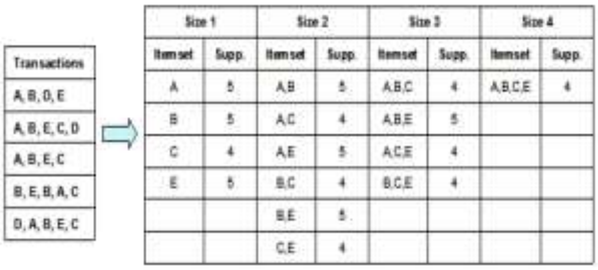
V. Pattern Analysis

The types and levels of analysis, performed on the integrated usage data, depend on the ultimate goals of the analyst and the desired outcomes. In this section we describe some of the most common types of pattern discovery and analysis techniques employed in the Web usage mining domain and discuss some of their applications.

Association rule discovery and statistical correlation analysis can find groups of items or pages that are commonly accessed or purchased together. This, in turn, enables Web sites to organize the site content more efficiently, or to provide effective cross-sale product recommendations.

Most common approaches to association discovery are based on the Apriori algorithm. This algorithm finds groups of items occurring frequently together in many transactions. Such groups of items are referred to as frequent itemsets. Association rules which satisfy a minimum confidence threshold are then generated from the frequent itemsets.

Recall an association rule is an expression of the form *X*→*Y* [*sup*, *conf*], where *X* and *Y* are itemsets, *sup* is the support of the itemset*X* ∪*Y* representing the probability that *X* and *Y* occurs together in a transaction, and *conf*is the confidence of the rule, defined by *sup*(*X*∪*Y*)/ *sup*(*X*), representing the conditional probability that *Y* occurs in a transaction given that *X* has occurred in that transaction.



Transactions
A, B, D, E
A, B, E, C, D
A, B, E, C
B, E, B, A, C
D, A, B, E, C

Size 1		Size 2		Size 3		Size 4	
Itemset	Supp.	Itemset	Supp.	Itemset	Supp.	Itemset	Supp.
A	5	A,B	5	A,B,C	4	A,B,C,E	4
B	5	A,C	4	A,B,E	5		
C	4	A,E	5	A,C,E	4		
E	5	B,C	4	B,C,E	4		
		B,E	5				
		C,E	4				

Table 1: Pattern Analysis In E-Commerce

VI. Consideration Of Data Using Patterns

The statistical analysis of pre-processed session data constitutes the most common form of analysis. In this case, data is aggregated by predetermined units such as days, sessions, visitors, or domains. Standard statistical techniques can be used on this data to gain knowledge about visitor behaviour. This is the approach taken by most commercial tools available for Web log analysis. Reports based on this type of analysis may include in-formation about most frequently accessed pages, average view time of a page, average length of a path through a site, common entry and exit points, and other aggregate measures. Despite a lack of depth in this type of analysis, the resulting knowledge can be potentially useful for improving the system performance, and providing support for marketing decisions. Furthermore, commercial Web analytics tools are increasingly incorporating a variety of data mining algorithms resulting in more sophisticated site and customer metrics.

Another form of analysis on integrated usage data is Online Analytical Processing (OLAP). OLAP provides a more integrated framework for analysis with a higher degree of flexibility. The data source for OLAP analysis is usually a multidimensional data warehouse which integrates us-age, content, and e-commerce data at different levels of aggregation for each dimension. OLAP tools allow changes in aggregation levels along each dimension during the analysis. Analysis dimensions in such a structure can be based on various fields available in the log files, and may include time duration, domain, requested resource, user agent, and referrers. This allows the analysis to be performed on portions of the log related to a specific time interval, or at a higher level of abstraction with respect to the URL path structure. The integration of e-commerce data in the data ware-house can further enhance the ability of OLAP tools to derive important business intelligence metrics [7]. The output from OLAP queries can also be used as the input for a variety of data mining or data visualization tools.

VII. General Access Patterns

Clustering is a data mining technique that groups together a set of items having similar characteristics. In the usage domain, there are two kinds of interesting clusters that can be discovered: user clusters and page clusters.

Clustering of user records are one of the most commonly used analysis tasks in Web usage mining and Web analytics. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in e-commerce applications or provide personalized Web content to the users with similar interests. Further analysis of user groups based on their demographic at-tributes (e.g., age, gender, income level, etc.) may lead to the discovery of valuable business intelligence. Usage-based clustering has also been used to create Web-based “user communities” reflecting similar interests of groups of users [5,4], and to learn user models that can be used to provide dynamic recommendations in Web personalization applications [3].

One straightforward approach in creating an aggregate view of each cluster is to compute the centroid (or the mean vector) of each cluster. The dimension value for each pageview in the mean vector is computed by finding the ratio of the sum of the pageview weights across transactions to the total number of transactions in the cluster. If pageview weights in the original transactions are binary, then the dimension value of a pageview*p* in a cluster centroid represents the percentage of transactions in the cluster in which *p* occurs. Thus, the centroid dimension value of *p* provides a measure of its significance in the cluster. Pageviews in the centroid can be sorted according to these weights and lower weight pageviews can be filtered out. The resulting set of pageview-weight pairs can be viewed as an “aggregate usage profile” representing the interests or behavior of a significant group of users.

More formally, given a transaction cluster *cl*, we can construct the ag-gregate profile pr_{cl} as a set of pageview-weight pairs by computing the centroid of *cl:*

$$pr_{cl} = \{(p, weight(p, pr_{cl})) \mid weight(p, pr_{cl}) \ge \mu\}, \qquad ...(1)$$

where:

- the significance weight, *weight*(*p*, pr_{cl}), of the page *p* within the aggre-gate profile pr_{cl} is given by

$$weight(p, pr_{cl}) = \frac{1}{|cl|}\sum_{s \in cl} w(p,s); \qquad ...(2)$$

| *cl*| is the number of transactions in cluster *cl*;

- *w*(*p*,s) is the weight of page *p* in transaction vector s of cluster *cl*; and
- the threshold $\mu$is used to focus only on those pages in the cluster that appear in a sufficient number of vectors in that cluster. Each such profile, in turn, can be represented as a vector in the original *n*-dimensional space of pageviews. This aggregate representation can be used directly for predictive modeling and in applications such as recommender systems: given a new user, *u*,who has accessed a set of pages, P_u, so far, we can measure the similarity of P_u to the discovered profiles, and recommend to the user those pages in matching profiles which have not yet been accessed by the user.

VIII. Customised Access Pattern

Classification is the task of mapping a data item into one of several predefined classes. In the Web domain, one is interested in developing a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of given the class or category. Classification can be done by using supervised learning algorithms such as decision trees, naive Bayesian classifiers, *k*-nearest neighbour classifiers, and Support Vector Machines. It is also possible to use previously discovered clusters and association rules for classification of new users [5].

Classification techniques play an important role in Web analytics applications for modeling the users according to various predefined metrics. For example, given a set of user transactions, the sum of purchases made by each user within a specified period of time can be computed. A classification model can then be built based on this enriched data in order to classify users into those who have a high propensity to buy and those who do not, taking into account features such as users' demographic attributes, as well their navigational activities.

Another important application of classification and prediction in the Web domain is that of collaborative filtering. Most collaborative filtering applications in existing recommender systems use *k*-nearest neighbour classifiers to predict user ratings or purchase propensity by measuring the correlations between a current (target) user's profile (which may be a set of item ratings or a set of items visited or purchased) and past user profiles in order to find users in the database with similar characteristics or preferences [6]. Many of the Web usage mining approaches discussed earlier can also be used to automatically discover user models and then apply these models to provide personalized content to an active user [386, 445].

Basically, collaborative filtering based on the *k*-Nearest-Neighbour (*k*NN) approach involves comparing the activity record for a target user with the historical records *T* of other users in order to find the top *k* users who have similar tastes or interests. The mapping of a visitor record to its neighbourhood could be based on similarity in ratings of items, access to similar content or pages, or purchase of similar items. In most typical col-laborative filtering applications, the user records or profiles are a set of ratings for a subset of items. The identified neighbourhood is then used to recommend items not already accessed or purchased by the active user. Thus, there are two primary phases in collaborative filtering: the neighbourhood formation phase and the recommendation phase. In the context of Web usage mining, *k*NN involves measuring the similarity or correlation between the target user's active session u (represented as a vector) and each past transaction vector v (where v ∈*T*). The top *k* most similar transactions to u are considered to be the neighbourhood for the session u. More specifically, the similarity between the target user, u, and a neighbour, v, can be calculated by the Pearson's correlation coefficient defined below:

$$sim(\mathbf{u},\mathbf{v}) = \frac{\sum_{i\in C}(r_{\mathbf{u},i}-\bar{r}_{\mathbf{u}})(r_{\mathbf{v},i}-\bar{r}_{\mathbf{v}})}{\sqrt{\sum_{i\in C}(r_{\mathbf{u},i}-\bar{r}_{\mathbf{u}})^2}\sqrt{\sum_{i\in C}(r_{\mathbf{v},i}-\bar{r}_{\mathbf{v}})^2}}, \qquad (5)$$

where *C* is the set of items that are co-rated by u and v (i.e., items that have been rated by both of them), $r_{u,i}$ and $r_{v,i}$ are the ratings (or weights) of some item *i* for the target user u and a possible neighbourv respectively, and u*r*andv*r*are the average ratings (or weights) of u and v respectively. Once similarities are calculated, the most similar users are selected.

IX. Conclusion

Web usage mining has emerged as the essential tool for realizing more personalized, user-friendly and business-optimal Web services[1,2]. Advances in data pre-processing, modeling, and mining techniques, applied to the Web data, have already resulted in many successful applications in adaptive information systems, personalization services, Web analytics tools, and content management systems. As the complexity of Web applications and user's interaction with these applications increases, the need for intelligent analysis of the Web usage data will also continue to grow.

Usage patterns discovered through Web usage mining are effective in capturing item-to-item and user-to-user relationships and similarities at the level of user sessions. However, without the benefit of deeper domain knowledge, such patterns provide little insight into the underlying reasons for which such items or users are grouped together. Furthermore, the inherent and increasing heterogeneity of the Web has required Web-based applications to more effectively integrate a variety of types of data across multiple channels and from different sources.

References

[1] Rekha Jain, Dr G.N.Purohit, “Page Ranking Algorithmsfor Web Mining,” International Journal of Computerapplication,Vol 13, Jan 2011.

[2] Cooley, R, Mobasher, B., Srivastava, J."Web Mining:Information and pattern discovery on the World Wide Web”.In proceedings of the 9th IEEE International Conference on tools with Artificial Intelligence (ICTAI’ 97).NewposrtBeach,CA 1997.

[3] Pooja Sharma, PawanBhadana, “Weighted Page ContentRank For Ordering Web Search Result”, International Journal of Engineering Science and Technology, Vol 2, 2010.

[4] R. Kosala, H. Blockeel “Web mining research” A survey.ACM Sigkdd Explorations,2(1):1-15, 2000.

[5] Wang jicheng, Huang Yuan,WuGangshan, Zhang Fuyan,"Web mining: Knowledge discovery on the Web Systems",Man and Cybernetics 1999 IEEE SMC 99 conference Proceedings. 1999 IEEE International conference

[6] Raymond Kosala, HendrikBlockee, "Web Mining Research: A Survey", ACM Sigkdd Explorations Newsletter, June2000, Volume 2.

[7] Taher H. Haveliwala, "Topic-Sensitive Page Rank: AContext-Sensitive Ranking Algorithms for Web Search", IEEE transactions on Knowledge and Data EngineeringVol.15, No 4 July/August 2003.