FRAMEWORK FOR WEB PERSONALIZATION USING WEB MINING

Monika Soni[1], Rahul Sharma[2], Vishal Shrivastava[3]

[1]*M. Tech. Scholar, Arya College of Engineering and IT, Rajasthan, India,* ***12.monika@gmail.com***
[2]*M. Tech. Scholar, B.M.S. College of Engineering, Punjab, India,* ***rahulmnu1@gmail.com***
[3]*Associate. Prof., Arya College of Engineering and IT, Rajasthan, India,* ***vishal500371@yahoo.co.in***

Abstract

WWW is a large amount of information provider and a very big source of information. Users are increasing every day for accessing web sites. For efficient and effective handling, web mining coupled with suggestion techniques provides personalized contents at the disposal of users. Web Mining is an area of Data Mining dealing with the extraction of interesting knowledge from the Web. Here we are presenting a comprehensive overview of the personalization process based on Web usage mining. In this a host of Web usage mining activities required for this process, including the pre-processing and integration of data from multiple sources, and common pattern discovery techniques that are applied to the integrated usage data.

Index Terms: *Web-Usage Mining, Data Mining, Personalization, Pattern Discovery, Web Mining, Web Personalization*

----------***----------

1. INTRODUCTION AND BACKGROUND

The tremendous growth in the number and the complexity of information resources and services on the Web has made Web personalization an indispensable tool for both Web-based organizations and for the end users. The ability of a site to engage visitors at a deeper level, and to successfully guide them to useful and pertinent information, is now viewed as one of the key factors in the site's ultimate success. Web personalization can be described as any action that makes the Web experience of a user customized to the user's taste or preferences. Principal elements of Web personalization include modelling of Web objects (such as pages or products) and subjects (such as users or customers), categorization of objects and subjects, matching between and across objects and/or subjects, and determination of the set of actions to be recommended for personalization.

There are several well-known drawbacks to content-based or rule-based filtering techniques for personalization. The type of input is often a subjective description of the users by the users themselves, and thus is prone to biases. The profiles are often static, obtained through user registration, and thus the system performance degrades over time as the profiles age. Furthermore, using content similarity alone may result in missing important "pragmatic" relationships among Web objects based on how they are accessed by users. Collaborative filtering [Herlocker et al., 1999; Konstan et al., 1997; Shardanand and Maes, 1995] has tried to address some of these issues, and, in fact, has become the predominant commercial approach in most successful e-commerce systems. These techniques generally involve matching the ratings of a current user for objects (e.g., movies or products) with those of similar users (nearest neighbours) in order to produce recommendations for objects not yet rated by the user. The primary technique used to accomplish this task is the k-Nearest-Neighbor (kNN) classification approach which compares a target user's record with the historical records of other users in order to find the top k users who have similar tastes or interests.

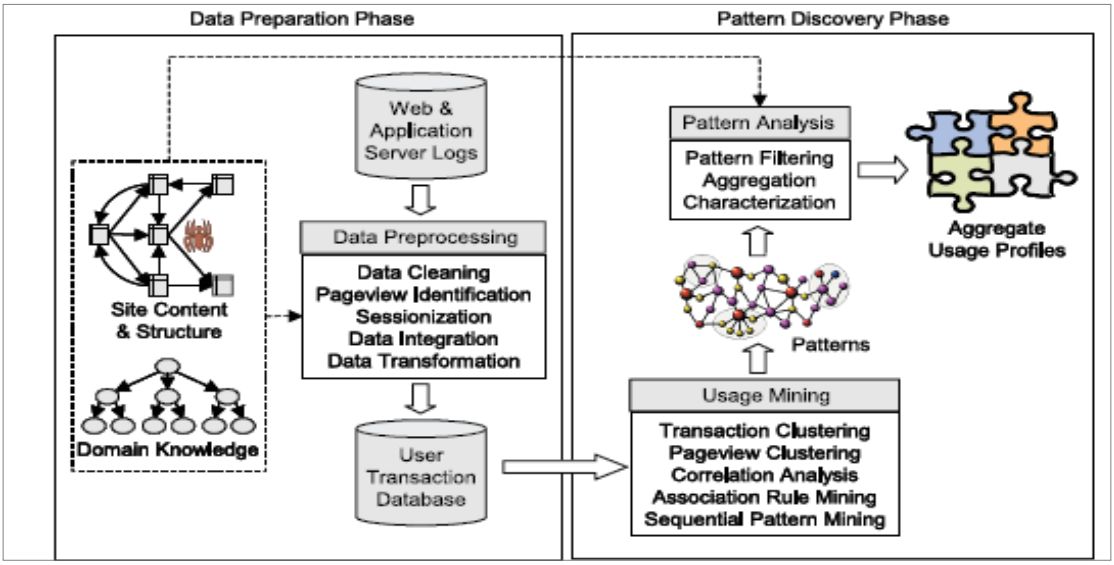


Figure 1:-The offline data preparation and pattern discovery components

.

2. DATA PREPARATION AND MODELLING

An important task in any data mining application is the creation of a suitable target data set to which data mining algorithms are applied. This 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. The data preparation process is often the most time consuming and computationally intensive step in the knowledge discovery process. Web usage mining is no exception: in fact, the data preparation process in Web usage mining, often requires the use of especial algorithms and heuristics not commonly employed in other domains. This process is critical to the successful extraction of useful patterns from the data. In this section we discuss some of the issues and concepts related to data modelling and preparation in Web usage mining.

While this discussion is in the general context of Web usage analysis, we are focused especially on the factors that have been shown to greatly affect the quality and usability of the discovered usage patterns for their application in Web personalization.

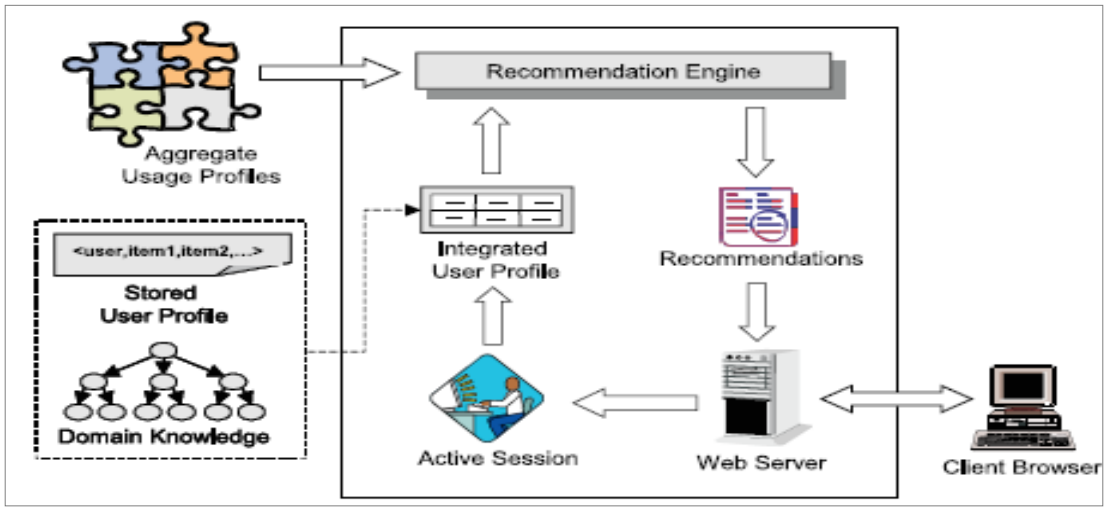


Figure 2:- Data Modelling

3. SOURCES AND TYPES OF DATA

The primary data sources used in Web usage mining are the server log files, which include Web server access logs and application server logs. Additional data sources that are also essential for both data preparation and pattern discovery include the site files and meta data, operational databases, application templates, and domain knowledge. Generally speaking, the data obtained through these sources can be categorized into four groups.

4. USAGE DATA

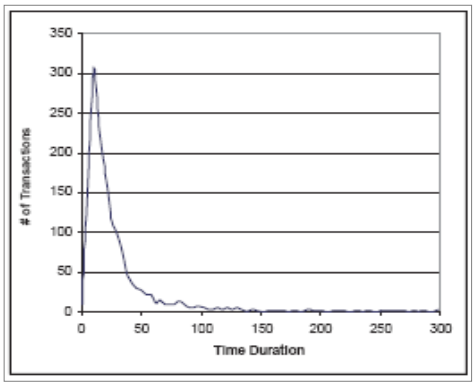
The log data collected automatically by the Web and application servers represents the fine-grained navigational behaviour of visitors. Depending on the goals of the analysis, this data needs to be transformed and aggregated at different levels of abstraction. In Web usage mining, the most basic level of data abstraction is that of a page view. Physically, a page view is an aggregate representation of a collection of Web objects contributing to the display on a user's browser resulting from a single user action (such as a click through). These Web objects may include multiple pages (such as in a frame-based site), images, embedded components, or script and database queries that populate portions of the displayed page (in dynamically generated sites). Conceptually, each page view represents a specific "type" of user activity on the site, e.g., reading a news article, browsing the results of a search query, viewing a product page, adding a product to the shopping cart, and so on. On the other hand, at the user level, the most basic level if behavioural abstraction is that of a server session (or simply a session). A session (also commonly referred to as a "visit") is a sequence of page views by a single user during a single visit. The notion of a session can be further abstracted by selecting a subset of page views in the session that are significant or relevant for the analysis tasks at hand. We shall refer to such a semantically meaningful subset of page views as a transaction (also referred to as an episode according to the W3C Web Characterization Activity [W3C]). It is important to note that a transaction does not refer simply to product purchases, but it can include a variety of types of user actions as captured by different pageviews in a session.

5. USAGE DATA PREPARATION

The required high-level tasks in usage data pre-processing include data cleaning, page view identification, user identification, session identification (or sessionization), the inference of missing references due to caching, and transaction (episode) identification. We provide a brief discussion of some of these tasks below; for a more detailed discussion see [Cooley, 2000; Cooley et al., 1999]. Data cleaning is usually site-specific, and involves tasks such as, removing extraneous references to embedded objects, graphics, or sound files, and removing references due to spider navigations. The latter task can be performed by maintaining a list of known spiders, and through heuristic identification of spiders and Web robots [Tan and Kumar, 2002]. It may also be necessary to merge log files from several Web and application servers. This may require global synchronization across these servers. In the absence of shared embedded session ids, heuristic methods based on the "referrer" field in server logs along with various sessionization and user identification methods (see below) can be used to perform the merging. Client- or proxy-side caching can often result in missing access references to those pages or objects that have been cached. Missing references due to caching can be heuristically inferred through path completion which relies on the knowledge of site structure and referrer information from server logs [Cooley et al., 1999]. In the case of dynamically generated pages, form-based applications using the HTTP POST method result in all or part of the user input parameter not being appended to the URL accessed by the user (though, in the latter case, it is possible to re-capture the user input through packet sniffers on the server side).

6. POST-PROCESSING OFUSER TRANSACTIONS DATA

In addition to the aforementioned pre-processing steps leading to user transaction matrix, there are a variety of transformation tasks that can be performed on the transaction data. Here, we highlight some of data transformation tasks that are likely to have an impact on the quality and action ability of the discovered patterns resulting from mining algorithms. Indeed, such post-processing transformations on session or transaction data have been shown to result in improvements in the accuracy of recommendations produced by personalization systems based on Web usage mining.



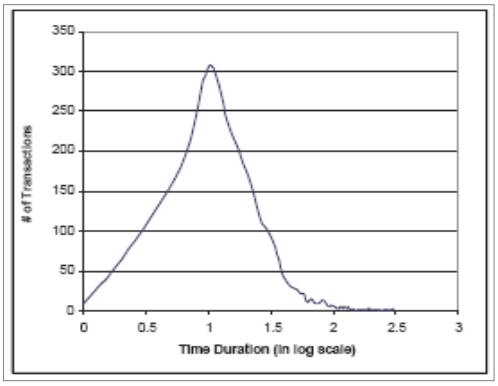


Figure 3:- Distribution of pageview durations: raw-time scale (left), log-time scale (right).

7. DATA INTEGRATION FROM MULTIPLE SOURCES

In order to provide the most effective framework for pattern discovery and analysis, data from a variety of sources must be integrated. Our earlier discussion already alluded to the necessity of considering the content and structure data in a variety of preprocessing tasks such as page view identification, sessionization, and the inference of missing data. The integration of content, structure, and user data in other phases of the Web usage mining and personalization processes may also be essential in providing the ability to further analyze and reason about the discovered patterns, derive more actionable knowledge, and create more effective personalization tools. For example, in e-commerce applications, the integration of both user data (e.g., demographics, ratings, purchase histories) and product attributes from operational databases is critical. Such data, used in conjunction with usage data, in the mining process can allow for the discovery of important business intelligence metrics such as customer conversion ratios and lifetime values. On the other hand, the integration of semantic knowledge from the site content or domain ontologies can be used by personalization systems to provide more useful recommendations. For instance, consider a hypothetical site containing information about movies which employs collaborative filtering on movie ratings or page view transactions to give recommendations. The integration of semantic knowledge about movies (possibly extracted from site content), can allow the system to recommend movies, not just based on similar ratings or navigation patterns, but also perhaps based on similarities in attributes such as movie genres or commonalities in casts or directors.

8. PATTERN DISCOVERY FROM WEB USAGE DATA

Levels and Types of Analysis

Different kinds of analysis can be performed on the integrated usage data at different levels of aggregation or abstraction. The types and levels of analysis, naturally, depend on the ultimate goals of the analyst and the desired outcomes. For instance, even without the benefit of an integrated e-commerce data mart, statistical analysis can be performed on the preprocessed session or transaction data. Indeed, static aggregation (reports) 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 (however, most such tools do not perform all of the necessary preprocessing tasks described earlier, this resulting in erroneous or misleading outcomes). Reports based on this type of analysis may include information 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 measure.

9. DATA MINING TASKS FOR WEB USAGE DATA

We now focus on specific data mining and pattern discovery tasks that are often employed when dealing with Web usage data. Our goal is not to give detailed descriptions of all applicable data mining techniques, but to provide some relevant background information and to illustrate how some of these techniques can be applied to Web usage data. In the next section, we present several approaches to leverage the discovered patterns for predictive Web usage mining applications such as personalization.

10. CLUSTERING APPROACHES

In general, there are two types of clustering that can be performed on usage transaction data: clustering the transactions (or users), themselves, or clustering pageviews. Each of these approaches are useful in different applications, and in particular, both approaches can be used for Web personalization. There has been a significant amount of work on the applications of clustering in Web usage mining, e-marketing, personalization, and collaborative filtering.

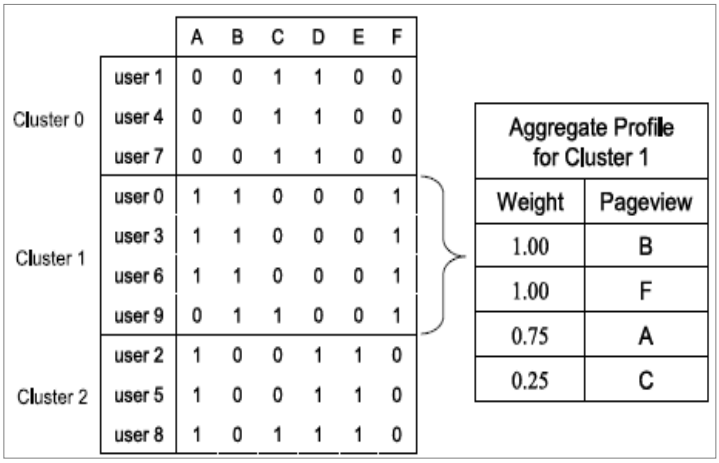


Figure 4:- deriving aggregate usage profiles from transaction clusters

For example, an algorithm called PageGather has been used to discover significant groups of pages based on user access patterns [Perkowitz and Etzioni, 1998]. This algorithm uses, as its basis, clustering of pages based the Clique (complete link) clustering technique. The resulting clusters are used to automatically synthesize alternative static index pages for a site, each reflecting possible interests of one user segment. Clustering of user rating records has also been used as a prior step to collaborative filtering in order to remedy the scalability problems of the k-nearest-neighbor algorithm [O'Conner and Herlocker, 1999]. Both transaction clustering and pageview clustering have been used as an integrated part of a Web personalization framework based on Web usage mining.

11. USING THE DISCOVERED PATTERNS FOR PERSONALIZATION

As noted in the Introduction section, the goal of the recommendation engine is to match the active user session with the aggregate profiles discovered through Web usage mining, and to recommend a set of objects to the user. We refer to the set of recommended object (represented by pageviews) as the recommendation set. In this section we explore the recommendation procedures to perform the matching between the discovered aggregate profiles and an active user's session. Specifically, we present several effective recommendation algorithms based on clustering (which can be seen as an extension of standard kNN-based collaborative filtering), association rule mining (AR), and sequential pattern (SP) or contiguous sequential pattern (CSP) discovery. In the cases of AR, SP, and CSP, we consider efficient and scalable data structures for storing frequent itemset and sequential patterns, as well as a recommendation generation algorithms that use these data structures to directly produce real-time recommendations (without the apriori generation of rule).

CONCLUSIONS AND OUTLOOK

In this chapter we have attempted to present a comprehensive view of the personalization process based on Web usage mining. The overall framework for this process was depicted in Figures 1.1 and 1.2. In the context of this framework, we have discussed a host of Web usage mining activities necessary for this process, including the preprocessing and integration of data from multiple sources, and pattern discovery techniques that are applied to the integrated usage data. We have also presented a number of specific recommendation algorithms for combining the discovered knowledge with the current status of a user's activity in a Web site to provide personalized content to a user. The approaches we have detailed show how pattern discovery techniques such as clustering, association rule mining, and sequential pattern discovery, performed on Web usage data, can be leveraged effectively as an integrated part of a Web personalization system.

REFERENCES

[1] R. Agarwal, C. Aggarwal, and V. Prasad. A Tree Projection Algorithm for Generation of Frequent Itemsets. In Proceedings of the High Performance Data Mining Workshop, Puerto Rico, April 1999.

[2] C. C. Aggarwal, J. L. Wolf, and P. S. Yu. A New Method for Similarity Indexing for Market Data. In Proceedings of the 1999 ACM SIGMOD Conference, Philadelphia, PA, June 1999.

[3] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94), Santiago, Chile, Sept 1994.

[4] R. Agrawal and R. Srikant. Mining Sequential Patterns. In Proceedings of the International Conference on Data Engineering (ICDE'95), Taipei, Taiwan, March 1995.

[5] A. Banerjee and J. Ghosh. Clickstream Clustering Using Weighted Longest Common Subsequences. In Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining, Chicago, Illinois, April 2001.

[6] B. Berendt, A. Hotho, and G. Stumme. Towards Semantic Web Mining. In Proceedings of the First International Semantic Web Conference (ISWC02), Sardinia, Italy, June 2002.

[7] B. Berendt, B. Mobasher, M. Nakagawa, and M. Spiliopoulou. The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis. In Proceedings of the 4th WebKDD 2002 Workshop, at the ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD'2000), Edmonton, Alberta, Canada, July 2002b.

[8] 7.B. Berendt and M. Spiliopoulou. Analysing Navigation Behaviour in Web Sites Integrating Multiple Information Systems. VLDB Journal, Special Issue on Databases and the Web, 9(1):56–75, 2000.

[9] J. Borges and M. Levene. Data Mining of User Navigation Patterns. In B. Masand and M. Spiliopoulou, editors, Web Usage Analysis and User Profiling: Proceedings of the WEBKDD'99 Workshop, LNAI 1836, pages 92–111. Springer-Verlag, 1999.

[10] A. Buchner and M. D. Mulvenna. Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining. SIGMOD Record, 4(27):54–61, 1999.

[11] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin. Combining Content-based and Collaborative Filters in an Online Newspaper. In Proceedings of the ACM SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation, Berkeley, California, August 1999.

[12] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns fromWeb Data. Ph. d. dissertation, Department of Computer Science, University of Minnesota, Minneapolis, Minnesota, 2000.

[13] R. Cooley, B. Mobasher, and J. Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns. Journal of Knowledge and Information Systems, 1(1):5–32, 1999.

[14] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to Construct Knowledge Bases from theWorldWideWeb. Artificial Intelligence, 118(1-2):69–113, 2000.

[15] H. Dai and B. Mobasher. Using Ontologies to Discover Domain-LevelWeb Usage Profiles. In Proceedings of the 2nd SemanticWeb Mining Workshop at ECML/PKDD 2002, Helsinki, Finland, August 2002.

[16] M. Deshpande and G. Karypis. Selective Markov Models for Predicting Web-Page Accesses. In Proceedings of the First International SIAM Conference on Data Mining, Chicago, April 2001.

[17] W. B. Frakes and R. Baeza-Yates. Information Retrieval: Data Structures and Algorithms. Prentice Hall, Englewood Cliffs, NJ, 1992.

BIOGRAPHIES:

Monika Soni Pursuing M. Tech. in Computer Science. She has published many national and international research papers. She has written 3 books for engineering and engineering diploma.

Rahul Sharma Pursuing M. Tech. in Computer Science & Engineering. He has published many national and international research papers. He has written 5 books for engineering and engineering diploma.

Vishal Shrivastava working as Assistant Professor in Arya College & IT He has published many national and international research papers. He has very depth knowledge of his research areas.