ISSN No. 0976-5697

Volume 5, No. 2, March 2014 (Special Issue)

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

An Application of Web Usage Mining Framework for Mining Dynamic Web Sites

G. Vijaya Kumari
SVECW, Bhimavaram.
gvijayakumari8@gmail.com

P.Praneeth
VKR,VNB&AGKEngineering College
pranith.145@gmail.com

V. Purushotham Raju
Assosiate Professor,
Shri Vishnu Engineering College for Women,
praju@svecw.edu.in

Abstract— In this paper, we present a complete framework and findings in mining Web usage patterns from Web log files of a real Website that has all the challenging aspects of real-life Web usage mining, including evolving user profiles and external data describing an ontology of the Web content. Even though the Web site under study is part of a nonprofit organization that does not "sell" any products, it was crucial to understand "who" the users were, "what" they looked at, and "how their interests changed with time," all of which are important questions in Customer Relationship Management (CRM). Hence, we present an approach for discovering and tracking evolving user profiles. We also describe how the discovered user profiles can be enriched with explicit information need that is inferred from search queries extracted from Web log data. Profiles are also enriched with other domain-specific information facets that give a panoramic view of the discovered mass usage modes. An objective validation strategy is also used to assess the quality of the mined profiles, in particular their adaptability in the face of evolving user behavior.

Index Terms—Mining evolving click streams, user profiles, Web usage mining, semantic web mining.

I. INTRODUCTION

Customer Relationship Management (CRM) can use data from within and outside an organization to allow an understanding of its customers on an individual basis or on a group basis such as by forming customer profiles. An improved understanding of the customer's habits, needs, and interests can allow the business to profit by, for instance, "cross selling" or selling items related to the ones that the customer wants to purchase. Hence, reliable knowledge about the customers' preferences and needs forms the basis for effective CRM. As businesses move online, the competition between businesses to keep the loyalty of their old customers and to lure new customers is even more important, since a competitor's Web site may be only one click away. The fast pace and large amounts of data available in these online settings have recently made it imperative to use automated data mining or knowledge discovery techniques to discover Web user profiles. These different modes of usage or the so-called mass user profiles can be discovered using Web usage mining techniques that can automatically extract frequent access patterns from the history of previous user clickstreams stored in Web log files. These profiles can later be harnessed toward personalizing the Web site to the user or to support targeted marketing.

Although there have been considerable advances in Web usage mining, there have been no detailed studies presenting a fully integrated approach to mine a real Web site with the challenging characteristics of today's Web sites, such as evolving profiles, dynamic content, and the availability of taxonomy or databases in addition to Web logs. In this paper, we present a complete framework and a summary of our experience in mining Web usage patterns with real-world challenges such as evolving access patterns, dynamic pages, and external data describing an ontology of the Web content and how it relates to the business actors (in the case of the studied Web site, the companies, contractors, consultants, etc., in corrosion). The Web site in this study is a portal that provides access to news, events, resources, company information (such as companies or contractors supplying related products and services), and a library of technical and regulatory documentation related to corrosion and surface treatment. The portal also offers a virtual meeting place between companies or organizations seeking information about other companies or organizations.

Without loss of generality, in the rest of this paper, we will refer to all the Web site participants (organizations, contractors, consultants, agencies, corporations, centers, agencies, etc.) simply as companies. The Web site in our study is managed by a nonprofit organization that does not sell anything but only provides free information that is ideally complete, accurate, and up to date. Hence, it was crucial to understand the different modes of usage and to know what kind of information the visitors seek and read on the Web site and how this information evolves with time. For this reason, we perform clustering of the user sessions extracted from the Web logs to partition the users into several homogeneous groups with similar activities and then extract user profiles from each cluster as a set of relevant URLs.

This procedure is repeated in subsequent new periods of Web logging (such as biweekly), then the previously discovered user profiles are tracked, and their evolution pattern is categorized.

II. AN OVERVIEW OF WEB USAGE MINING

Recently, data mining techniques have been applied to extract usage patterns from Web log data. This process, known as Web usage mining, is traditionally performed in several stages to achieve its goals:

a. Collection of Web data such as activities/clickstreams recorded in Web server logs.

b. Preprocessing of Web data such as filtering crawlers requests, requests to graphics, and identifying unique sessions,

c. Analysis of Web data, also known as Web Usage Mining, to discover interesting usage patterns or profiles, and

b. Interpretation/evaluation of the discovered profiles.

In this paper, we further added a fifth step after a repetitive application of steps 1-4 on multiple time periods, i.e.,

e. Tracking the evolution of the discovered profiles.

A. Integrating Semantics in Web Usage Mining:

Relying only on Web usage data for user modeling or for personalization can be inefficient, either when there is insufficient usage data for the purpose of mining certain patterns or when new pages are added and thus do not accumulate sufficient usage data at first. The lack of usage data in these cases can be compensated by adding other information such as the content of Web pages or the structure of a Web site. In the keywords that appear in Web pages are used to generate document vectors, which are later clustered in the document space to further augment user profiles. The Website's own hierarchical structure is treated like an implicit taxonomy or concept hierarchy that is exploited in computing the similarity between any two Web pages on the Web site.

This allows a better comparison between sessions that contain visits to Web pages that are different and yet semantically related (for example, under the same more general topic). The idea of exploiting concept hierarchies or taxonomies has already been found to enhance association rule mining and to facilitate information searching in textual data. Even though keywords that are present in the Web pages have been used to add a "content" aspect to usage data, the keyword-based approach remains incapable of capturing more complex relationships at a deeper semantic level. Thus, a general framework was proposed for using domain ontologies to automatically characterize usage profiles containing a set of structured Web objects.

The advent of dynamic URLs mostly in tandem with Web databases has recently made it even more difficult to interpret URLs in terms of user behavior, interests, and intentions. For instance, consider the following cryptic association rule within the context of an online bookstore, "If http://www.the_shop.com/show.html-?item=123, then http://www.the_shop.com/show.html?item=456, support = 0.05, and confidence = 0.4." A more meaningful rule would be "users who bought Hamlet also tended to buy How to Stop Worrying and Start Living." This, in turn, has motivated which mined patterns of application events instead of patterns of URLs by exploiting the semantics of the visited pages. Within this spirit, Service-based concept hierarchies were introduced earlier for analyzing the search behavior of visitors, that is, "how they navigate rather than what they retrieve." In this case, concept hierarchies form the basic method of grouping Web pages together before Web usage mining. Usage mining was enhanced by describing the user behavior in terms of an ontology underlying a particular Web site.

The semantic annotation of the Web content was assumed to have been performed a priori, since the Web site in question was a knowledge portal with an inherent RDF annotation. In order to mine interesting patterns, first, the Web logs were semantically enriched with ontology concepts. Then, these semantic Web logs were mined to extract patterns such as groups of users, users' preferences, and rules. Following a similar approach, Web usage logs were enriched with semantics derived from the content of the Web site's pages. Content keywords were first mapped to the categories of a manually constructed domain-specific taxonomy through the use of a thesaurus, and then the Web documents were clustered based on the taxonomy categories. The enhanced Web logs, called C-Logs, were then used as input to Web usage mining.

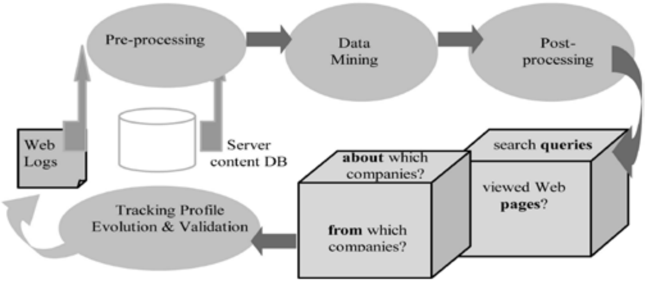


Figure.1. Web usage mining process and discovered profile facets.

The automatic identification of user profiles is a knowledge discovery task consisting of periodically mining new contents of the user access log files and is summarized in the following steps:

a. Preprocess Web log file to extract user sessions.

b. Cluster the user sessions by using Hierarchical Unsupervised Niche Clustering (H-UNC).

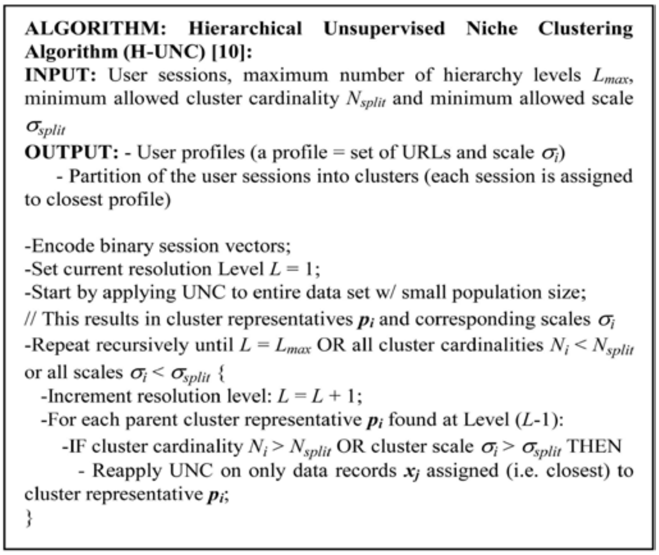
c. Summarize session clusters/categories into user profiles.

d. Enrich the user profiles with additional facets by using additional Web log data and external domain knowledge.

e. Track current profiles against existing profiles.

To cluster user sessions, we use H-UNC [10], a divisive hierarchical version of a robust clustering approach (Unsupervised Niche Clustering (UNC)) that uses a Genetic Algorithm (GA) to evolve a population of candidate solutions through generations of competition and reproduction.

Table1



ALGORITHM: Hierarchical Unsupervised Niche Clustering Algorithm (H-UNC) [10]:

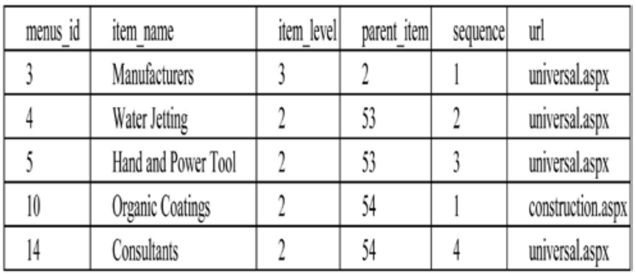
INPUT: User sessions, maximum number of hierarchy levels L_{max}, minimum allowed cluster cardinality N_{split} and minimum allowed scale σ_{split}

OUTPUT: - User profiles (a profile = set of URLs and scale σ_i)
- Partition of the user sessions into clusters (each session is assigned to closest profile)

```
-Encode binary session vectors;
-Set current resolution Level $L = 1$;
-Start by applying UNC to entire data set w/ small population size;
// This results in cluster representatives $\boldsymbol{p_i}$ and corresponding scales $\sigma_i$
-Repeat recursively until $L = L_{max}$ OR all cluster cardinalities $N_i < N_{split}$ or all scales $\sigma_i < \sigma_{split}$ {
  -Increment resolution level: $L = L + 1$;
  -For each parent cluster representative $\boldsymbol{p_i}$ found at Level ($L$-1):
     -IF cluster cardinality $N_i > N_{split}$ OR cluster scale $\sigma_i > \sigma_{split}$ THEN
        - Reapply UNC on only data records $\boldsymbol{x_j}$ assigned (i.e. closest) to cluster representative $\boldsymbol{p_i}$;
}
```

Partial Taxonomy of a Few Dynamic URLs (Identified by Base URL (url) and Parameter (menus_id))

Table: 2



menus_id	item_name	item_level	parent_item	sequence	url
3	Manufacturers	3	2	1	universal.aspx
4	Water Jetting	2	53	2	universal.aspx
5	Hand and Power Tool	2	53	3	universal.aspx
10	Organic Coatings	2	54	1	construction.aspx
14	Consultants	2	54	4	universal.aspx

III. ONCLUSION

We presented a framework for mining, tracking, and validating evolving multifaceted user profiles on Web sites that have all the challenging aspects of real-life Web usage mining, including evolving user profiles and access patterns, dynamic Web pages, and external data describing ontology of the Web content. A multifaceted user profile summarizes a group of users with similar access activities and consists of their viewed pages, search engine queries, and inquiring and inquired companies.

IV. REFERENCES

[1]. R. Cooley, B. Mobasher, and J. Srivastava, “Web Mining: Information and Pattern Discovery on the World Wide Web,” Proc. Ninth IEEE Int’l Conf. Tools with AI (ICTAI ’97), pp. 558-567, 1997.

[2]. O. Nasraoui, R. Krishnapuram, and A. Joshi, “Mining Web Access Logs Using a Relational Clustering Algorithm Based on a Robust Estimator,” Proc. Eighth Int’l World Wide Web Conf. (WWW ’99), pp. 40-41, 1999.

[3]. O. Nasraoui, R. Krishnapuram, H. Frigui, and A. Joshi, “Extracting Web User Profiles Using Relational Competitive Fuzzy Clustering,” Int’l J. Artificial Intelligence Tools, vol. 9, no. 4, pp. 509-526, 2000.