Asian Journal of Information Technology 10 (2): 78-83, 2011
ISSN: 1682-3915

Extracting Web User Profiles Using H-UNC Clustering

S. Karthikeyan and M. Hakkeem
Department of Computer Science and Engineering,
SNS College of Technology, Coimbatore-35, India

Abstract: Web usage mining applies data mining techniques to records of Web site visits. To better understand patterns of usage, analysis should take the semantics of visited URLs into account. 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, researchers 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.

Key words: User profiles, web usage mining, semantic web mining, CRM, queries, profile

INTRODUCTION

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.

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 click streams 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. In this study, we present a complete framework and a summary of the 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 study, we will refer to all the Web site participants (organizations, contractors, consultants, agencies, corporations, centers, agencies, etc.) simply as companies.

AN OVERVIEW OF WEB USAGE MINING

Recently, data mining techniques have been applied to extract usage patterns from Web log data (Cooley *et al*., 1997; Nasraoui *et al*., 1999; Nasraoui *et al*., 2000; Srivastava *et al*., 2000; Spiliopoulou and Faulstich, 1998). This process known as Web usage mining is traditionally performed in several stages (Cooley *et al*., 1997; Nasraoui and Krishnapuram, 2000) to achieve its goals:

- Collection of Web data such as activities/click streams recorded in Web server logs
- Preprocessing of Web data such as filtering crawlers requests, requests to graphics and identifying unique sessions

Corresponding Author: S. Karthikeyan, Department of Computer Science and Engineering,
SNS College of Technology Coimbatore-35, India

- Analysis of Web data also known as Web Usage Mining (Srivastava *et al*., 2000) to discover interesting usage patterns or profiles
- Interpretation/evaluation of the discovered profiles
- Tracking the evolution of the discovered profiles

Web usage mining can use various data mining or machine learning techniques to model and understand Web user activity. In Nasraoui and Krishnapuram (2002), clustering was used to segment user sessions into clusters or profiles that can later form the basis for personalization. Nasraoui *et al*. (2003) proposed the notion of an adaptive Web site was proposed where the user's access pattern can be used to automatically synthesize index pages. The research of Cooley *et al*. (1997) is based on using association rule discovery as the basis for modeling Web user activity whereas the approach proposed by Maloof and Michalski (1995) used probabilistic grammars to model Web navigation patterns for the purpose of prediction.

The approach in Maloof and Michalski (2000) proposed building data cubes from Web log data and later applying Online Analytical Processing (OLAP) and data mining on the cube model. New fuzzy relational clustering techniques were used to discover user profiles that can overlap (Srivastava *et al*., 2000) whereas robust clustering (Nasraoui *et al*., 2000) was proposed to mine profiles that are resistant to noise that is naturally present in click stream data. A robust density based evolutionary clustering technique was proposed to discover an optimal number of multi resolution and robust user profiles (Mitchell *et al*., 1994).

Handling profile evolution: Most previous research efforts in Web usage mining have worked with the assumption that the Web usage data is static. However, the dynamic aspects of Web usage have recently become important (Fig. 1). This is because Web access patterns on a Web site are dynamic due not only to the dynamics of Web site content and structure but also to changes in the user's interests and thus their navigation patterns. Maloof and Michalski (2000) classified online learning in the presence of concept drift as either evolutionary or revolutionary with regard to adaptation to change. An evolutionary scheme modifies existing knowledge based on completely new training examples (for example, STAGGER (Schlimmer and Granger, 1986) whereas a revolutionary approach discards old knowledge and learns new knowledge from the new training examples (for example, window-based techniques (Widmer and Kubat, 1996). A third approach includes hybrids that inherit from both the revolutionary and evolutionary approaches.

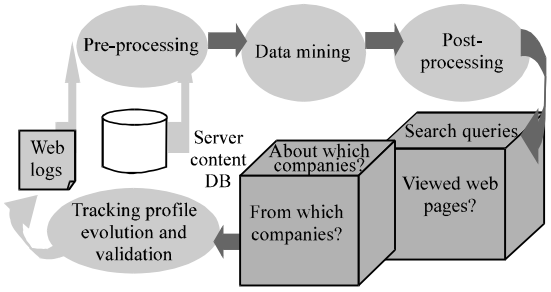

Fig. 1: Web usage mining process and discovered profile facets

For instance, Mitchell's Calendar Learning Apprentice (Mitchell *et al*., 1994) learns new decision rules from training data and incorporates these new rules into the existing knowledge base. Maloof and Michalski (1995) further classified the way online learning systems work into three different modes: no memory, partial memory or full memory. In the no memory mode, the system does not use any past training examples for updating the current model (for example, STAGGER (Schlimmer and Granger, 1986) whereas in the partial memory mode, a subset of the previously seen training examples is used for later learning. Finally, in the full memory mode, all past training examples are used in updating an existing model. A continuum between no memory and full memory (gradual forgetting) used a forgetting function-based approach in supervised learning (Koychev, 2000) and to cluster evolving streams (Nasraoui *et al*., 2003).

PROFILE DISCOVERY BASED ON WEB USAGE MINING

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:

- Preprocess Web log file to extract user sessions
- Cluster the user sessions by using Hierarchical Unsupervised Niche Clustering (H-UNC) (Nasraoui and Krishnapuram, 2002)
- Summarize session clusters/categories into user profiles
- Enrich the user profiles with additional facets by using additional Web log data and external domain knowledge
- Track current profiles against existing profiles

Preprocessing the web log file to extract user sessions: The access log of a Web server is a record of all files (URLs) accessed by users on a Web site. Each log entry consists of the access time, IP address, URL viewed, REFERRER (the Web page visited just prior to the current one), etc.

The first step in preprocessing (Cooley *et al*., 1997; Nasraoui *et al*., 1999) consists of mapping the N_u URLs on a Web site to distinct indices. A user session consists of requests from the same IP address within a predefined time period. Each URL in the site is assigned a unique number jε1, . . .N_u,

Where:

N_u = The total number of valid URLs
ith = The user session is then encoded
N_u = The dimensional binary attribute vector
$s^{(i)}$ = The following property

$$S_j^{(i)} = \{1 \text{ if user I accessed URL}_j\}; \{0 \text{ otherwise}\}$$

Clustering sessions into an optimal number of categories: To cluster user sessions, we use H-UNC (Mitchell *et al*., 1994), a divisive hierarchical version of a robust clustering approach (Unsupervised Niche Clustering (UNC) (Nasraoui and Krishnapuram, 2000) that uses a Genetic Algorithm (GA) (Holland, 1975) to evolve a population of candidate solutions through generations of competition and reproduction.

The main outline of the H-UNC algorithm is sketched in the following. The reason that we use H-UNC Instead of other clustering algorithms is that unlike most other algorithms, H-UNC can handle noise in the data and automatically determines the number of clusters. In addition, evolutionary optimization allows the use of any domain-specific optimization criterion and any similarity measure in particular a subjective measure that exploits domain knowledge or ontologies.

However, unlike purely evolutionary search-based algorithms, H-UNC combines evolution with local Piccard updates to estimate the scale σ_i of each profile, thus converging fast (about 20 generations). H-UNC is outlined as follows (more details can be found in (Mitchell *et al*., 1994).

Algorithm (Hierarchical unsupervised niche clustering algorithm (H-UNC)) (Mitchell *et al*., 1994)

Input: User sessions, maximum number of hierachy 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 p_i and corresponding scales σ_i
Repeat recursively unit L = L_{max} OR all cluster cardinalities $N_i \leq N_{split}$ or all scales $\sigma_i < \sigma_{split}$ {
Increment resolution level: L = L+1;
For each parent cluster representative 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 x_j assigned (i.e., closest) to cluster representative p_i;

Post processing and enrichment of session clusters: After automatically grouping sessions into different clusters, we summarize the session categories in terms of user profile vectors pi (Nasraoui *et al*., 2000; Srivastava *et al*., 2000). The kth component/weight of this vector (pi_k) captures the relevance of URL_k in the ith profile as estimated by the conditional probability that URL_k is accessed in a session belonging to the ith cluster (this is the frequency with which URL_k was accessed in the sessions belonging to the ith cluster). Web pages, the profile properties include the following facets:

Search queries: These are queries submitted to search engines before visiting the Web site for sessions that belong to this profile.

Inquiring companies: These are companies/organizations of registered users or unregistered users whose IP addresses can be mapped.

Inquired companies: These are companies/organizations that have been inquired about during the sessions belonging to this profile.

Enriching user profiles with search query terms (Search Queries): In addition to the relevant URLs that are extracted from the sessions assigned to each profile, reseaarchers can extract information about the explicit information need of the users in each profile from the queries that they could have typed prior to visiting the Web site when this information is available from the readily available REFERRER field in the Web log files. Hence, for each profile, we accumulate all the search phrases extracted from the REFERRER fields of the assigned user sessions. This allows us to describe each profile in terms of either a set of significant URLs or a set of explicit search query phrases and terms.

Enriching user profiles with inquiring company Information (from companies): In addition to the relevant URLs that are extracted from the sessions assigned to each profile, we can extract information about which companies or organizations tend to visit the Web site and

fall in this profile. We extract this information from two complementary sources: by getting the company information that corresponds to an ID in the server content Database where the ID is extracted from the Web log file in case the visitors register and sign in through the registration page or if the visitors did not sign in through the registration page, then an attempt is made to obtain the company affiliation from a specialized Web service (www.whois.com).

This can be queried with an IP address via an API to determine not only what information was found relevant on the Web site but also to whom it was relevant to help support further personalization efforts.

Enriching user profiles with queried company Information (about companies): The Web site under study provides a virtual meeting point between different companies providing various services that are related to the portal's subject. Hence, it was important to know not only which companies take part in each cluster of activities but also what company information seemed to be relevant to users in each cluster. For this reason, in addition to the relevant URLs that are extracted from the sessions assigned to each profile, we extracted information about which companies have been inquired about by visitors in this profile in case a user searches and clicks on one of the listed companies contact information on the Web site. We parse the identity of the company from the Web log file and map it to a specific company via the server content database.

EXPLOITING AN EXTERNAL ONTOLOGY FOR MAPPING AND RELATING DYNAMIC WEB PAGES

Most of today's Web sites deliver a large number of URLs if not only dynamic URLs. A dynamic URL is a page address hat results from the search of a database driven Web site or Web site that runs a script. Unlike static URLs in which the contents of the Web page do not change, dynamic URLs are typically generated from specific queries to a site's database.

Even though the examples given in the following discussion consistently use the ASP extension, this extension can be replaced by any other dynamic URL extension (such as PHP), without any changes in the generic approach. Although, static Web pages tend to have meaningful URLs such as /reports/fall_2003/ benefits.html, most dynamic URLs such as/universal. aspx?id = 55 and codes_id = 60 are unfortunately hard to discern or even recognize based only on their URL.

We resolved this issue by resorting to available external data 2 that maps database contents to a dynamic resource and its parameter values. The ASP codes in most

Table 1: Partial taxonomy of a few dynamic URLs (identified by base URL and parameter (menus_id))

Menus_id	Item_name	Item_ level	Parent_ item	Sequence	URL
3	Manufactures	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

Table 2: Taxonomy data for the dynamic URL universal.aspx? id = 56

Menus_id	Item_name	Item level	Parent_ item	Sequence	URL
56	Regulations and laws	1	4939	1	Universal.aspx
4939	NST center and reg;	0	-	1	NST

menus can be mapped during the preprocessing phase to a parent/child structure by using external data (Table 1) thus mapping URLs to meaningful hierarchical descriptions. Insertion is done in reverse order from the end to the start of the final composed label until we reach the parent at level 0. Both implicit (URL itself) and explicit (Table 1) taxonomy information are seamlessly incorporated into the session clustering via the computation of the special session similarity measure (Table 2).

TRACKING EVOLVING USER PROFILES

Tracking different profile events across different time periods can generate a better understanding of the evolution of user access patterns and seasonality. Note that both profiles and click streams are typically evolving, since the profiles are nothing more than summaries of the click streams which are themselves evolving. Each profile pi is discovered along with an automatically determined measure of scale σi that represents the amount of variance or dispersion of the user sessions in a given cluster around the cluster representative.

This measure is used to determine the boundary around each cluster (an area located at a distance σi from the profile pi) and thus allows us to automatically determine whether two profiles are compatible. About 2 profiles are compatible if their boundaries overlap. The notion of compatibility between profiles is essential for tracking evolving profiles.

After mining the Web log of a iven period, we perform an automated comparison between all the profiles discovered in the current batch and the profiles discovered in the previous batch by a sequence of SQL queries on the profiles that have been stored in a database, as shown in the Track Profiles Algorithm. A typical query for retrieving corresponding profiles between Periods T_i and T_{i+1} is SELECT This Profile, To this Profile FROM Profile Trail WHERE Period = T_i (Fig. 2).

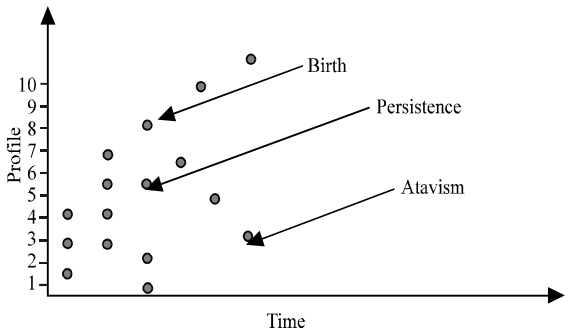

Fig. 2: Visualization of the profile evolution

Algorithm (TrackProfiles)
Input: Discovered profiles for all Time Periods stored in Database //(profile = set of relevant URLs and scale σ_i)
Beginning time period T_i ending time period T_k

Output (Profile Trail): Profile-to-Profile tracking Table from Time Period T_i to Time Period T_k//

For I = Time Period T_i to Time Period T_k do
For J = first profile in Time Period 1 to last profile in Time Period 1 do
For K = first profile in Time Period 1+1 to last profile in Time Period 1+1 Do {
Distance [k] = S_{web} (profile, $profile_k$);
IF Distance [k] <σ_i THEN insert into profileTrail (Period, ThisProfile, TothisProfile) values (1, Profile [J], K);

ASYSTEMATIC APPROACH TO PROFILE AND EVOLUTION VALIDATION

As a summary, profiles represent a reduced form of the data that is at the same time as close as possible to the original input data. This description is reminiscent of an information retrieval scenario in the sense that profiles that are retrieved should be as close as possible to the original session data. Closeness should take both of the following into account.

Precision: A summary profile's items are all correct or included in the original input data that is they include only the true data items.

Coverage/recall: A summary profile's items are complete compared to the data that is summarized that is they include all the data items. These criteria are clearly contradictory, since precision will favor only the smallest profiles, eventually with a single URL whereas coverage will favor the largest possible profiles. Ideally, each data query should be answered by a profile that is identical to this query. However, this is unrealistic, since it requires the profile's summary to be identical to the entire input database. Therefore, the summary should consist of the smallest number of profiles that are as similar as possible to the input data. The validation procedure (Nasraoui and Goswami, 2006) attempts to answer the following questions:

- Is the data set completely summarized/represented by the mined profiles/patterns?
- Is the data set faithfully/accurately summarized/ represented by the mined profiles/patterns?

CONCLUSION

Researchers 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.

REFERENCES

Cooley, R., B. Mobasher and J. Srivatsava, 1997. Web mining: Information and pattern discovery on the world wide web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, Nov. 3-8 Newport Beach, CA., pp: 558-567.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems. 1st Edn., MIT Press, Cambridge, Mass.

Koychev, I., 2000. Gradual forgetting for adaptation to concept drift. Proceedings of the ECAI Workshop Current Issues in Spatio-Temporal Reasoning, (ECAI'00), Berlin, pp: 101-106.

Maloof, M.A. and R.S. Michalski, 1995. Learning evolving concepts using partial memory approach. Proceedings of the AAAI Fall Symposium on Active Learning, Nov. 10-12, Cambridge, MA, pp: 10-12.

Maloof, M.A. and R.S. Michalski, 2000. Selecting examples for partial memory learning. Machine Learning, 41: 27-52.

Mitchell, T., R. Garuana, D. Freitag, J. McDermott and D. Zabowski, 1994. Experience with a learning personal assistant. Commun. ACM, 37: 81-91.

Nasraoui, O. and R. Krishnapuram, 2000. A novel approach to unsupervised robust clustering using genetic niching. Proceedings of the 9th IEEE International Conference Fuzzy Systems, May 7-10, Memphis State University Press, pp: 170-177.

Nasraoui, O. and R. Krishnapuram, 2002. A new evolutionary approach to web usage and context sensitive associations mining. Int. J. Computat. Intelli. Applic., 2: 339-348.

Nasraoui, O. and S. Goswami, 2006. Mining and validating localized frequent itemsets with dynamic tolerance. Proceedings of the 6th SIAM International Conference Data Mining, April 20-22, Maryland, pp: 578-582.

Nasraoui, O., C. Cardona, C. Rojas and F. Gonzalez, 2003. Mining evolving user profiles in noisy web clickstream data with a scalable immune system clustering algorithm. Proceedings of the 5th Workshop Web Mining as a Premise to Effective and Intelligent Web Applications, Aug. 27, USA., pp: 71-81.

Nasraoui, O., H. Frigui, R. Krishnapuram and A. Joshi, 2000. Extracting web user profiles using relational competitive fuzzy clustering. Int. J. Artificial Intelli. Tools, 9: 509-526.

Nasraoui, O., R. Krishnapuram and A. Joshi, 1999. Mining web access logs using a relational clustering algorithm based on a robust estimator. Proceedings of the 8th International World Wide Web Conference, May 11-14, Canada, pp: 40-41.

Schlimmer, J.C. and R.S. Granger, 1986. Incremental learning from noisy data. Machine Learning, 1: 317-357.

Spiliopoulou, M. and L.C. Faulstich, 1998. WUM: A web utilization miner. Proceedings of the 1st International Workshop on Web and Databases (WDB'98), Spain, pp: 241-253.

Srivastava, J., R. Cooley, M. Deshpande and P.N. Tan, 2000. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorat., 1: 1-12.

Widmer, G. and M. Kubat, 1996. Learning in the presence of concept drift and hidden contexts. Machine Learning, 23: 69-101.