Mining and Tracking Evolving User Profiles and More – A Real Life Case Study

Olfa Nasraoui, Maha Soliman, and Antonio Badia
{olfa.nasraoui,mmsoli01,anbadia01@louisville.edu}
Dept. of Computer Science and Engineering
Speed School of Engineering
University of Louisville
Louisville, KY, 40291

ABSTRACT

Personalization tailors a user's interaction with the Web information space based on information gathered about them. Declarative user information such as manually entered profiles continue to raise privacy concerns and are neither scalable nor flexible in the face of very active dynamic Web sites and changing user trends and interests. One way to deal with this problem is through a completely automated Web personalization system. Such a system can be based on Web usage mining to discover Web usage profiles, followed by a recommendation system that can respond to the users' individual interests. While there have been considerable advances in the field of Web usage mining, there have been no detailed case studies presenting fully integrated approaches to mine a real website with the challenging characteristics of today's websites, such as *evolving* access patterns and *dynamic* content. We present a case study summarizing our preliminary approach and findings in mining web usage patterns from the Web log files of a real life website that has all the challenging aspects of real life web usage mining, including evolving user profiles and access patterns, dynamic web pages, and external data describing an ontology of the web content. We also present a simple approach to enrich the discovered user profiles with *explicit information need* as inferred from search queries extracted from the Web log data.

Keywords
Web personalization, web recommendation, web usage mining, collaborative filtering, mining evolving clickstreams.

1. INTRODUCTION

Customer Relationship Management or CRM is a collection of business methods that aim at understanding the customers of an enterprise, in order to improve the performance of the organization [31]. CRM uses information from data sources within and outside an organization to allow understanding of its customers, either on an individual or 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 the customer wants to purchase, as well as "upselling" or selling item that the customer wants, but in a more expensive model or with additional options. 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 website may in fact be only one click away. However, in the online setting, the available wealth of data about the customers and potential customers is not only more varied and easier to collect (for example clickstreams that can indicate the interests of a Web user), but it is also available in large quantities, and can immediately be put to use as part of a real-time CRM strategy. These online considerations have recently made it imperative to use automated and reliable data mining or knowledge discovery techniques to discover Web user profiles. The ability to understand the different modes of usage or so called mass user profiles on a heavily visited website is a major asset. This understanding can be accomplished thanks to a variety of Web usage mining techniques [1,2,3,4,5,6,9,15-20] that can automatically extract frequent access patterns from the history of previous user clickstreams stored in web log files. While there have been considerable advances in the field of Web usage

mining, there have been no detailed case studies presenting fully integrated approaches to mine a real website with the challenging characteristics of today's websites, such as *evolving access patterns* and *dynamic content*. We present a case study summarizing our preliminary approach and findings in mining web usage patterns from the Web log files of a real life website that has all the challenging aspects of real life web usage mining, including evolving user profiles and access patterns, dynamic web pages, and external data describing an ontology of the web content. We also present a simple approach to enrich the discovered user profiles with *explicit information need* as inferred from search queries extracted from the Web log data. Because of space limitations, we do not discuss other pertinent details such as cleaning the Web logs from the massive requests that originate from the various Web crawlers or bots. The website in this study is a portal offered by the National Surface Treatment (*NST*) Center which partners with the Navy, DoD operations and industry to fight corrosion and solve coating problems. The NSTCenter Web site (www.nstcenter.com) is currently considered as one of the main information repositories for the corrosion-treatment community, providing access news, events, resources, and an extensive library of technical and regulatory documentation related to corrosion and surface treatment.

The framework for our web usage mining and a roadmap to the rest of the paper is shown in Figure 1, which starts with the collection of Web server logs, follows with standard pre-processing, such as data cleaning and sessionization, then continues with the pattern discovery through clustering, and ends with the implementation of a recommendation strategy for personalizing a website to a variety of users.

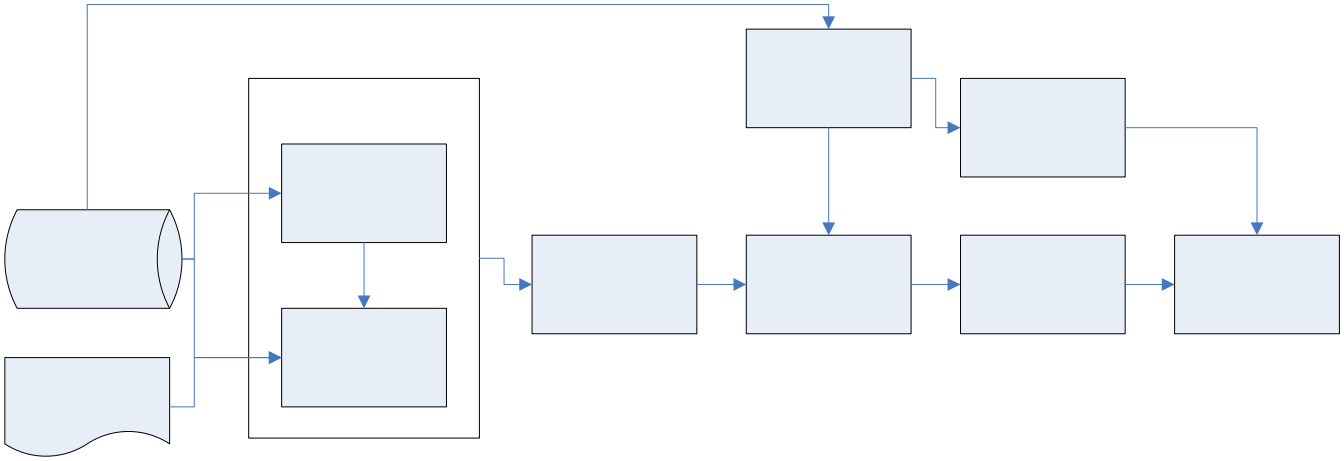

Figure 1: The life cycle of web usage mining in the presented study and a roadmap for the paper

The rest of the paper is organized as follows. In Section 2, we present an overview of our approach to profile discovery using Web usage mining. In Section 3, we discuss our approach to handle dynamic content and exploit external data describing an ontology of the web content. In Section 4, we discuss our approach and present our results in tracking evolving user profiles. In Section 5, we discuss some possible recommendation strategies based on the discovered profiles. We also illustrate a possible approach to enrich Web user profiles with explicit Information Need extracted from the REFERRER information in Web logs, and present a recommendation strategy based on automated "Information Need" assessment. Finally, in Section 6, we present our conclusions.

2. PROFILE DISCOVERY BASED ON WEB USAGE MINING

The first step in intelligent profile-based Web personalization is the automatic identification of user profiles [1,2,3,4,5,6,9,15-20]. This constitutes the *knowledge discovery engine*. These profiles are later used to recommend relevant URLs to old and new anonymous users of a Web site. This constitutes the *recommendation engine*

[9,15,23,24,25,27]. The knowledge discovery part can be executed *offline* by periodically mining *new* contents of the user access log files, and can be summarized in the following steps:

(1) Preprocess log file to extract user *sessions*,
(2) *Categorize* sessions by *Hierarchical Unsupervised Niche Clustering (H-UNC)* [7,8,9]
(3) Summarize the session categories in terms of *user profiles*,
(4) Infer *context-sensitive* URL *associations* from user profiles,

Step 1: 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 following information components: *access time, IP address, URL viewed, REFERRER (*web page visited *just prior to* visiting the website under study)*…etc*. An example showing two entries is displayed below, where the bold items are the requested URL and the underlined items are the search query terms for a request that originated from clicking on one of the results of user's enquiry on a search engine.

```
2004-04-04 14:18:00 x.y.z.w - W3SVC1 NT-NSTC a.b.c.d 80 GET /universal.asp id=55&codes_id= 200 0 0 733 31
www.nstcenter.com Mozilla/4.0+(compatible;+MSIE+6.0;+MSNIA;+Windows+98;+AT&T+CSM6.0;+T312461)
http://www.nstcenter.com/company.asp
2004-04-20 20:43:42 x.y.z.w - W3SVC1 NT-NSTC a.b.c.d 80 GET /nstc_009_pdfs/009-32_FY05.pdf - 200 64 0 491 344
www.nstcenter.com Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+4.0) http://www.google.com/search?hl=en&ie=UTF-8&oe=UTF-
8&q=Sigma+Glaze+5492&btnG=Google+Search
```

The first entry shows a request from IP address `x.y.z.w` (encoded to guard the user's privacy in this example) to a dynamic URL: ***/universal.asp*** with parameters *id* assigned a value of 55 and *codes_id* assigned a NULL value. The REFERRER field in this case shows that the web page visited immediately before the current URL was `http://www.nstcenter.com/company.asp`.

The second entry shows a request from IP address `x.y.z.w` (encoded to guard the user's privacy in this example) to a static URL: **`/nstc_009_pdfs/009-32_FY05.pdf`.** The REFERRER field in this case shows that the web page visited immediately before the current URL was the results page of a search on Google:

`http://www.google.com/search?hl=en&ie=UTF-8&oe=UTF-8&q=Sigma+Glaze+5492&btnG=Google+Search,`

where we have underlined the terms of the search query ("*Sigma Glaze*"). Whenever possible, we save the referrer search query terms for sessions, to be added to the profiles discovered by Web usage mining, as explained later in section 5.

The first step in preprocessing [1,2,3] consists of mapping the N_U URLs on a website to distinct indices. A user session consists of accesses originating from the same IP address within a predefined time period. Each URL in the site is assigned a unique number $j \in 1, \ldots, N_U$, where N_U is the total number of valid URLs. Thus, the i^{th} user session is encoded as an N_U-dimensional binary attribute vector $s^{(i)}$ with the property

$$s_j^{(i)} = \begin{cases} 1 & \text{if user accessed } j^{th} \text{ URL} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In addition to the URLs in each session, we encode the search query terms in case the REFERRER indicates a previous search.

Step 2: Clustering Sessions into an Optimal Number of Categories

For this task, we use *Hierarchical Unsupervised Niche Clustering* [6] or *H-UNC***.** *H-UNC* is a *hierarchical* version of a *robust genetic* clustering approach (*UNC*) [5], inspired by nature. A Genetic Algorithm (GA) [10] evolves a population of candidate solutions through generations of competition and reproduction until convergence to *one* solution. Hence, the GA cannot maintain population *diversity*. *Niching* methods, on the other

hand, attempt to maintain a *diverse* population with members distributed among *niches* corresponding to multiple solutions. An initial population of randomly selected sessions is encoded into binary chromosome strings that compete based on a density based fitness measure that is highest at the centers of good (dense) clusters. Different niches in the fitness landscape correspond to distinct clusters in the data set. The main outlines of the H-UNC algorithm are sketched below. Note that the clusters *and their number* are determined automatically, and that noise and outliers are tolerated. More details on H-UNC can be found in [6].

Hierarchical Unsupervised Niche Clustering Algorithm (*H-UNC*):

INPUT: user sessions, minimum allowed cluster cardinality and scale

OUTPUT: User profiles & a partition of the user sessions into clusters (each session is assigned to closed profile)

```
-Encode binary session vectors
-Set current resolution Level L = 1
-Start by applying UNC to entire data set w/ small population size;
-Repeat recursively until cluster cardinality or scale become too small {
   -Increment resolution level: L = L + 1
   -For each parent cluster found at Level (L-1):
       -Reapply UNC [5] only on data subset assigned to this parent
     cluster
       -Extract more child clusters at higher resolution (L > 1)
 }
```

Step 3: Summarizing Session Clusters into User Profiles

After automatically grouping sessions into different clusters, we summarize the session categories in terms of *user profile vectors* [3,4], $\boldsymbol{p}_i$**:** The k^{th} component/weight of this vector ($\boldsymbol{p}_{ik}$) captures the *relevance* of URL_k in the i^{th} profile, as estimated by the conditional probability that URL_k is accessed in a session belonging to the i^{th} cluster (this is the frequency with which URL_k was accessed in the sessions belonging to the i^{th} cluster). The model is further extended to a *robust profile* [3,6] based on robust weights that assign *only* sessions with high robust weight to a cluster's *core*. Unpolluted by noisy (irrelevant) sessions, these profiles give a *cleaner* description of the user interests. In addition to the notion of robust profiles, each profile $\boldsymbol{p}_i$ is discovered along with an automatically determined measure of scale σ_i that represents the amount of variance or dispersion of the user sessions in a given profile around the profile representative. This measure will later serve a very important role in determining the boundary around each cluster, and thus allow us to automatically determine whether two profiles are compatible or not. The notion of compatibility between profiles is essential for tracking evolving user profiles, as described later in section 4. A simple recommendation strategy based on assigning new users to the closest profile is discussed in Section 5.1.

Step 4: Enriching User Profiles with Search Query Terms

In addition to the relevant URLs which are extracted from the sessions assigned to each profile, we 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 our website, 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 either in terms of a set of significant URLs as was done in Step 3 above, or as a set of explicit search query phrases and terms. An intuitive recommendation strategy and a method to bridge query terms to clickstreams are discussed in Sections 5.2 and 5.3.

3. EXPLOITING AN EXTERNAL ONTOLOGY FOR MAPPING AND RELATING DYNAMIC WEB PAGES

Most of today's websites deliver a large number of, if not only dynamic web pages. While static web pages tend to have meaningful URLs such as `/reports/fall_2003/benefits.html`, most dynamic URLs, such as `/universal.asp id=55&codes_id=60`, are unfortunately hard to discern or even recognize based only on

their URL. This is because most dynamic web pages are identified based on special codes or values assigned to the parameters of a CGI program, and that are later used to pull specific content from linked databases. In order to resolve this issue, we have resorted to available external data, provided by the website designers, that maps the different database contents to a specific dynamic program and a set of specific parameter values. The ASP codes in the majority of the menus can be mapped to a hierarchical like structure by using external data such as shown in the following table (Table 1) that is exploited in the pre-processing phase by mapping the dynamic URLs to *hierarchically related and more meaningful descriptions*.

Table 1: A Taxonomy of Dynamic URLs (identified by base URL (*url*) and parameter (*menus_id*))

menus_id	item_name	item_level	parent_ite	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
7	Other Methods and Techniques	2	53	5	universal.aspx
10	Organic Coatings	2	54	1	construction.aspx
11	Inorganic Coatings	2	54	2	construction.aspx
14	Consultants	2	54	4	universal.aspx
15	Contractors	2	54	5	universal.aspx
22	Inspection and QA	2	55	9	universal.aspx
23	Cathodic Protection	2	55	1	universal.aspx
32	Training	2	57	7	universal.aspx
37	Corrosion Control	2	63	5	universal.aspx
40	Life Cycle Cost Considerations	2	58	5	universal.aspx
44	Reports	2	59	7	universal.aspx
45	Coatings	2	61	1	universal.aspx
46	Surface Preparation	2	61	2	construction.aspx
54	Coatings	1	4941	2	construction.aspx
55	Manufacturers / Suppliers	1	4942	1	construction.aspx
57	QA and Inspections	1	4941	3	universal.aspx

We illustrate our mapping procedure below with a list of options on the main menu clickable on the main page of http://www.nstcenter.com/ (left frame):

Regulations and Laws In the Log, this is only recorded as: `universal.aspx?id=56`. Table 2 below lists its content information (*Regulations and Law*). Furthermore, its parent (at level 0) is item code 4939 with label *NST Center&reg*. Hence, this URL is mapped to a "semantic label"**:** *NST Center&reg/Regulations and Laws*.

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

menus_id	item_name	item_leve	parent_ite	sequence	url
56	Regulations and Laws	1	4939	1	universal.aspx
4939	NST Center®	0		1	nst

Note, the last row above lists "`nst`" as one of the URLs and identifies it in the "uppermost" level (0) of the taxonomy, which is not a URL, but rather a *general area of the website*. Table 3 lists the remaining general areas with level 0:

Table 3: Taxonomy Data for the dynamic URL: `universal.aspx?id=56`.

menus_id	item_name	item_leve	parent_ite	sequence	url
4939	NST Center®	0		1	nst
4940	Navy Community	0		2	navy
4941	Surface Treatment	0		3	surface
4942	Company Connection	0		4	company

Air Quality and Emission Standards In the Log, this is only recorded as: `universal.aspx?id=6770`. In Table 4, this URL is mapped to a "semantic label": *Air Quality and Emission Standards* and has `parent-item =56`. In pre-processing, to exploit this hierarchical relationship via our *structure/hierarchy sensitive URL similarity in (3)*, we will replace this URL by `label[parent-item]/label[item]` which is "*NST Center&reg/Regulations and Laws/Air Quality and Emission Standards*"

Table 4

menus_id	item_name	item_level	parent_item	sequence	url
6770	Air Quality and Emission Standards	2	56	1	universal.aspx

In general, we need to read the parent of each item, and then recursively map a dynamic URL such as `universal.aspx?id=6770` to a string consisting of tokens separated by "/" where tokens are labels[parent-items]. Insertion is done in reverse order from the end to start of the final "composed" label until we reach the parent at level 0 (where we stop). First, we read from Table 1 all parent relation pairs and levels of items: The steps for this mapping are outlined in the following pseudocode:

```
Algorithm MapDynamicURL
Input: Dynamic_URL //(e.g. universal.aspx?id=6770), TaxonomyTable (e.g. Table 1-4)
Output: Final_label // (e.g. NST Center&reg/Regulations and Laws/Air Quality and Emission Standards)
Read <item.code, item.label, item.level, item.parent_item> from <TaxonomyTable>
Initialize parent = Dynamic_URL.current-item; // this is the code immediately following
"universal.aspx?id=" or "construction.aspx?id=" or any "url.aspx" where url is any of the urls
listed in the last column of Table 1

Initialize Final_label = ""; // initially blank label
WHILE parent.level >= 0 DO{
    Final_label = parent.label + "/" + Final_label; // concatenate parent & child labels
    parent = parent.parent_item;
}
```

A site map is shown in Table 5, listing some of the dynamic menu items and sub-items that are clickable from the website left frame. Figure 2 shows a small portion of the same website, but with the dynamic URLs untranslated.

3.1 Relation to Content-Based Filtering

Our approach incorporates information about the web pages content. However, this is slightly different from methods based on explicit content of the web pages. Instead, we infer this information from external knowledge. Also this knowledge is of a hierarchical nature, since it is in the form of a *taxonomy*. Rather than a pure content-based filtering approach [25], we use content only for the purpose of forming collaborative profiles, and therefore our approach falls in the class referred to by Pazzani in [24] as "*collaborative via content*" personalization. In "*collaborative via content*" personalization, content is used only to compute the similarity between a user profile and other user profiles, and not the similarity between a user profile and candidate items. Our approach can also be considered as a hybrid personalization strategy [27] since it merges different sources of data (*usage*, *content* through an external ontology, and *structure* through the incorporation of website structure in the similarity function in Equation (3)). Our approach to integrate semantics has been previously explored, but mostly in the field of content-based filtering and rule based recommender systems [26,28,29,30].

Table 5: Partial Site Map (most of the links are accessible through dynamic URLs, many items were removed because of space limitations as indicated by (…))

Home	**Technical Resources**	**Company Connection**
About Us	Surface Treatment	Manufacturers / Suppliers
Events	Surface Preparation	Cathodic Protection
News	Abrasive Blasting	(…)
Navy Community	(…)	Coatings
Coatings Approval Process (RoadMap)	Water Jetting	Automotive
Approved Exterior Ship Coatings	(…)	Marine
Approved Interior Ship Coatings	Hand and Power Tool	Containment and Control
Approved Exterior Submarine Coatings	Supplemental Resources	Dehumidification
Exterior Non-Skid	Chemical Removal	Equipment
Approved Interior Submarine Coatings	(…)	Standards
Sanitary Tanks	Environmental Considerations	(…)
Ballast Tanks	Supplemental Resources	
Auxiliary Tanks	Coatings	Robotic Systems
Potable Water Tank	Organic Coatings	Safety
Bilges	Epoxy Coatings	Soda Blasting
Technical Needs and Commercial Solutions	Alkyds	Surface Preparation
	(…)	Abrasive Blasting

U.S. Navy Freeboard & Superstructure Coatings
U.S. Navy Anticondensation Coating Program
U.S. Navy Heat Resistant Coating Program
U.S. Navy High Durability Coatings
U.S. Navy High Solids Tank Coating Program
(…)
Navy Preservation Working Groups
NAVSEA/Fleet Coating Working Groups
Underwater Hull
Nonskid
Tanks & Voids
Tanks and Voids Final Teleconference
Topside
Mega Rust Working Groups
Preservation Executive Oversight
Contracting & QA
Membership
Action Items
Paint Warranty
Membership
Action Items
Training
Membership
Action Items
Information Systems
Other Topics and Links
Shipyards
SUPSHIPs and Regional Maintenance Centers

Inorganic Coatings
High Temperature
(…)
QA and Inspections
Training
Companies
Navy Paint Inspector Training
Safety Considerations
(…)
Conference Proceedings
DOD
GSA SuperStore
Navy
Naval Ships Technical Manual (NSTM)
NSTM: Nonskid Extracts
NAVSEA Approved PPIs
(…)
Presentations
Reports
Standards and Specifications
ANSI
(…)
Regulations and Laws
Air Quality and Emission Standards
(…)
Publications
Books
Journals
Magazines
Dictionary

(…)
Walnut Shell Blasting
Consultants
Coatings
Marine
Surface Preparation
Water Jetting
Abrasive Blasting
(…)
Societies and Councils
Contractors
Coating
Marine
Abrasive Blasting
(…)
Testing and Evaluation
Abrasive Blasting
Water Jetting
(…)
Environmental
Current Report
Report Archive
Environmental - July 2005
(…)
EPA
OSHA
Air Quality
Coating Disposal

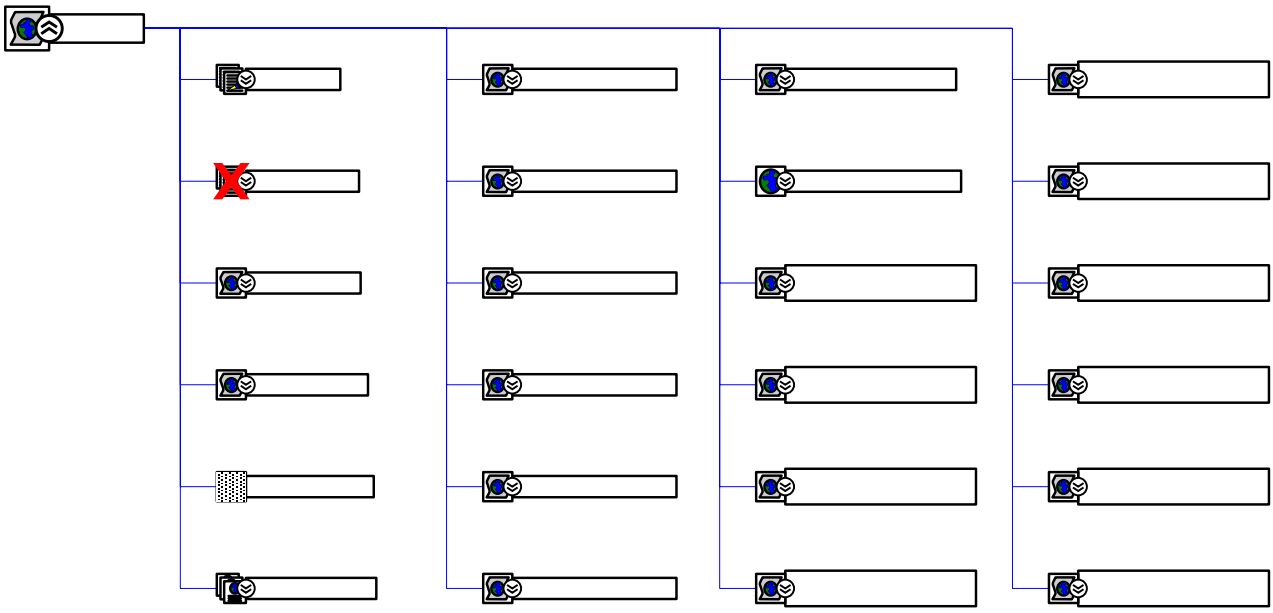

Figure 2: A portion of the website showing dynamic URLs

4. TRACKING EVOLVNG USER PROFILES

4.1 Mining User profiles from Anonymous Web Usage Data

Hierarchical Unsupervised Niche Clustering (H-UNC) [6] was applied on a set of web sessions preprocessed from the Web log data of the NSTC Web site. After filtering out irrelevant entries, the data was segmented into unique sessions based on the client IP address and a timeout threshold: The maximum elapsed time between two consecutive accesses in the same session was set to 45 minutes. We applied H-UNC, as part of the pattern discovery phase as shown in Figure 1, which illustrates the entire life cycle of the web profile mining process. HUNC was applied to the Web sessions using a maximal number of levels, L = 10, in the hierarchy, and the following parameters that control the final resolution: N_{split} = 30, and σ_{split} = 0.01. H-UNC partitioned the Web users sessions into several clusters at level 10, and each cluster was characterized by one of the profile vectors, $\boldsymbol{p}_i$. Some of these profiles are summarized in Table 6. The first profile is visualized in Figure 3 by projecting some of its relevant URLs on the website and highlighting the hyperlinks linking to them from previous pages.

TABLE 6. Examples of Discovered Profiles' URLs (with relevance weight per URL) and typical search queries extracted from the REFERRER information (Underlined and bold items indicate information inferred from the website ontology of what would otherwise appear as encoded dynamic URLs. This information is exploited by the Web Session Similarity given by (3), that takes into account the website hierarchical structure and/or relations between different dynamic URLs based on the website ontology)

weight	Profile	Some Search Queries
1.00	/Top_Frame.Aspx	"Soda Blasting"
1.00	/Menu.Aspx/Menu_File=Nst.Js	"Ballast Tank"
1.00	/Left1.Htm	"Navy Corrosion"
1.00	/Nst.Js	
1.00	/Frames.Aspx/Menu_File=Nst.Js&Main=/Dictionary2.Aspx	
1.00	/Mainpage_Frameset.Aspx/Menu_File=Nst.Js&Main=/Dictionary2.Aspx	
0.14	/Dictionary2.Aspx/Start_Letter=S	
0.64	/Construction.Aspx/Id=**NST Center&reg /Technical Library**	
0.64	/Mainpage_Frameset.Aspx/Menu_File=Nst.Js&Main=Universal.Aspx	
0.14	/Universal.Aspx/Id= **Nst Center&Reg;/ Technical Library/ Navy**	
0.21	/Dictionary2.Aspx/Start_Letter=M	
0.28	/Construction.Aspx/Id=**NST Center&reg/Publications**	
0.21	/Dictionary2.Aspx/Start_Letter=D	
0.21	/Universal.Aspx/Id=**Nst Center® /Technical Library/Corrosion Primer**	
0.14	/Dictionary2.Aspx/Start_Letter=I	
0.14	/Mainpage_Frameset.Aspx/Menu_File=Nst.Js&Main=**Ordnance Painting Systems**	
1.00	/Frames.Aspx/Menu_File=.Js&Main=/Universal.Aspx	"Shell Blasting"
0.98	/Top_Frame.Aspx	"Gavlon Industries
0.94	/Left1.Htm	"Obrien Paints"
0.10	/Favico.Ico	"Induron Coatings"
1.00	/Mainpage_Frameset.Aspx/Menu_File=.Js&Main=/Universal.Aspx	"Epoxy Polyamide"
1.00	/Menu.Aspx/Menu_File=.Js	
0.12	/Universal.Aspx/Id=**Company Connection/Manufacturers / Suppliers/Soda Blasting**	

4.2 Tracking Evolving Access Patterns

Tracking different profile events across different time periods can generate better understanding of the *evolution of user access patterns and seasonality*, and can be used to generate a simple visualization of the evolution of user trends on a website. Each profile $\boldsymbol{p}_i$ is discovered along with an automatically determined measure of scale σ_i that represents the *amount of variance or dispersion of the user sessions in a given profile around the profile representative*. This measure can be used to determine the *boundary* around each cluster which is an area located at a distance $=\sigma_i$ from the profile $\boldsymbol{p}_i$, and thus allow us to automatically determine whether two profiles are compatible. Two profiles are compatible if and only if each one of them is within the other's boundary. *The notion of compatibility between profiles is essential for tracking evolving user profiles*. After mining the web log entries from each batch corresponding to a given period, we perform an automated comparison between all the profiles discovered in the current batch and the profiles discovered in the previous batch, and determine which new profiles are compatible with the old profiles and which new profiles are incompatible with any previous profile. These last two cases give rise to two kinds of events: *Persistence* and *Birth*, respectively. A third event, *Death*, arises in case an old profile finds no compatible profile from the new batch. It is also possible to track profile *re-emergence* in the long term. This is the case of an old profile that disappears, and then reappears again when it is found to be compatible with a brand new profile in the current batch. This event is labeled as *Atavism*. We can plot the temporal dynamics of profiles birth, persistence, death, and atavism (rebirth) by labeling the x-axis with the periods corresponding to the different web log batches that undergo web usage mining: period1, period 2, …, etc; while the y-axis is used to indicate the profile index: new profiles are expanded vertically by adding new indices on top of existing ones. Finally we generate a plot depicting the website user trend evolution by adding a special symbol whenever profile (y) appears in period (x), and possibly adding event labels, such as Birth, Death, and Atavism, as these occur. This idea is illustrated in Figure 4. Note that this tracking is performed offline and takes advantage of a database management system to accelerate the access to archived user profiles and the comparison process.

Table 7 illustrates the results of the automated profile tracking and comparison process from the months of June 2004 up to September 2004 (the other months are omitted because of lack of space). Some months were too large for processing in one batch; therefore they were divided into two halves. For these months such as August, we use Aug 2004-Part1, Aug 2004-Part2 to indicate the first half and second half of August 2004, respectively.

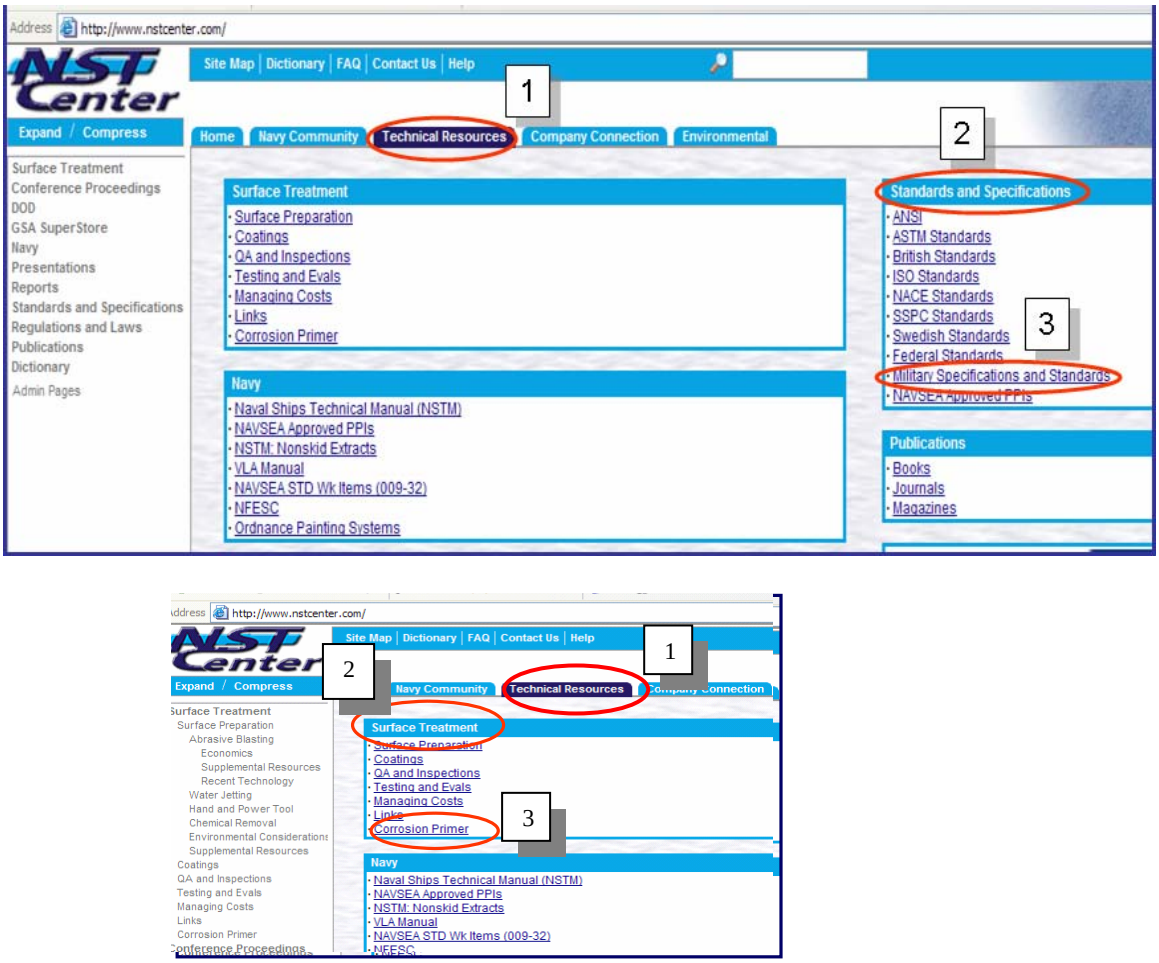

Figure 3: Visualizing the First Profile from Table 6 by Highlighting Relevant URLs

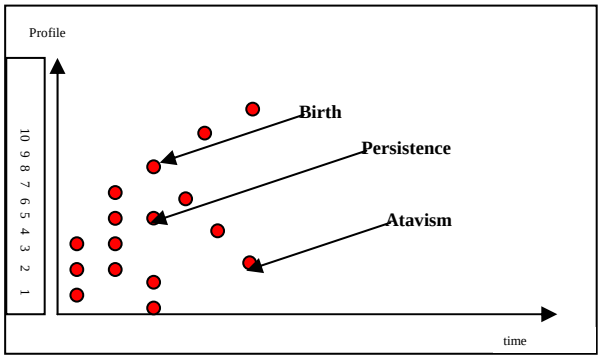

Figure 4: Visualization of the Profile Evolution

5. USING THE DISCOVERED PROFILES FOR CLICKSTREAM BASED AND SEARCH QUERY BASED RECOMMENDATIONS

Let $U = \{\text{url}_1, \text{url}_2, \ldots, \text{url}_{N_U}\}$ be a set of N_U urls on a given web site visited in web user sessions $\mathbf{s}_j$, $j = 1, \ldots, N_s$, as defined in (1). Let $P = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{N_P}\}$ be the set of N_P Web user profiles computed by the profile discovery engine. Each profile consists of a set of URLs associated with their relevance weights in that profile, and can be

viewed as a relevance vector of length N_U, with p_{ik} = relevance of url_k in the i^{th} profile. The problem of recommendation can be stated as follows. Given a current Web user session vector, $\boldsymbol{s}_j = [s_{j1}, s_{j2}, \ldots, s_{jN_U}]$, predict the set of URLs that are most relevant according to the user's interest, and recommend them to the user, usually as a set of *links* dynamically appended to the contents of the Web document returned in response to the most recent Web query. Because the degree of relevance of the URLs that are determined of interest to the user, may vary, it may also be useful to associate the k^{th} recommended URL with a corresponding URL relevance *score*, r_{jk}. Hence it is practical to denote the recommendations for current Web user session, $\boldsymbol{s}_j$, by a vector $\boldsymbol{r}_j = [r_{j1}, r_{j2}, \ldots, r_{jN_U}]$. In this study, we limit the scores to be binary.

5.1 Nearest Profile Prediction Based Recommender System

We adopt a collaborative filtering approach based on a set of anonymous mass user profiles, instead of individual user profiles or ratings [9,15,23,24,25]. Since we do take some content information (a taxonomy from external sources and not the web pages themselves) to compute the similarity between users, our approach is further classified as a "*collaborative via content*" [24]. The approach presented in Section 5.2 and 5.3 differs from previous approaches in that more information is used about the current user and past users than is typically stored in profiles. This includes the "*Information Need*" or "*Intent*" as inferred from the search query terms extracted from the *REFERRER* field. This information is not a predicate of the items in the user profile, but rather can be considered as a type of in-context behavioral attribute. Regardless of the sources of information that we incorporate in our profile building, the overall approach is still based on social or collaborative filtering. When all the above sources of information are combined, our personalization approach can be best described as a hybrid approach [27], and more specifically as "*collaborative via content and intention*".

The simplest and most rudimentary approach to Web recommendation is to simply determine the most similar profile to the current session, and to recommend the URLs in this profile, together with their URL relevance weights as URL recommendation scores.

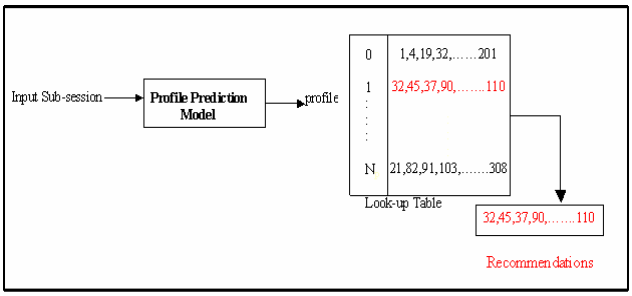

Figure 5: Nearest profile based Recommender System.

Figure 5 shows the structure of such a recommendation system, where the profile prediction model simply consists of a nearest-profile estimator based on computing a session to profile similarity, and selecting the profile with highest similarity as the predicted profile.

The similarity score between an input session, $\boldsymbol{s}$, and the i^{th} profile, $\boldsymbol{p}_i$, can be computed using the cosine similarity as follows,

$$S_{si}^{\cos ine} = \frac{\sum_{k=1}^{N_U} p_{ik} s_k}{\sqrt{\sum_{k=1}^{N_U} p_{ik} \sum_{k=1}^{N_U} s_k}} \qquad (2)$$

If a hierarchical Web site structure should be taken into account, then a modification of the cosine similarity, introduced in [3,4], that can take into account the Website structure can be used to yield the following similarity measure,

$$S_{si}^{web} = \max\left\{ \frac{\sum_{l=1}^{N_U}\sum_{k=1}^{N_U} p_{il} S_u(l,k) s_k}{\sum_{k=1}^{N_U} p_{ik} \sum_{k=1}^{N_U} s_k}, S_{si}^{\cos ine} \right\} \qquad (3)$$

where $\boldsymbol{S_u}$ is a URL to URL similarity matrix that is computed based on the amount of overlap between the paths leading from the root of the website (main page) to any two URLs, and is given by

$$S_u(i,j) = \min\left(1, \frac{|p_i \cap p_j|}{\max\left(1, \max\left(|p_i|, |p_j|\right) - 1\right)}\right) \quad (4)$$

We refer to the special similarity in (3) as the *Web Session Similarity*. *This web similarity takes into account not only the hierarchical structure of website content as inferred from the URL address itself, but also, possibly how different content items on the website relate to each other according to an externally defined website ontology*. In this case, we use a simple ontology based on '*is a*' relationships (i.e. a *taxonomy*) between individual dynamic URLs and higher level categories encoded in a website ontology. This similarity is used in our clustering algorithm (HUNC) [6] to group similar user sessions into clusters or profiles.

5.2 Bridging Pure Browsers and Pre-meditated Searchers

To estimate the user's "*initial information need*" or purpose of visiting the website, we need to consider the REFERRER field for the very first URL clicked in each session, where the REFERRER includes a search query (on a search engine such as *Netscape*, *Google*, *MSN* or *Yahoo*). Later on, after the user profiles are discovered, the frequency of the search query terms are tallied incrementally to characterize each profile. This can be used to label even the sessions in the profile that do not come from a search engine (pure browsing based sessions) with search terms of similar sessions (i.e. in the same profile) that did get initiated from an explicit search, and hence help bridge the gap between

1. *Pure Browsers:* users who just visit this website without coming from a search engine, and
2. *Premeditated Searchers:* users who visit as a result of clicking on one of the results of a query submitted to a search engine.

This is one way to enrich our knowledge of the user's Information Need. In fact, the Information Need assessment goes both ways: knowing one, we can predict the other, i.e.

1. From *Pure Browsers' clicks or URLs*, we can predict the most likely search terms that would describe the *Information Need* for this user
2. From the *search query terms* that would describe the Information Need of a Premeditated Searcher (i.e. from the first URL clicked into our website), predict the most likely *pure browsing clicks* that would be initiated by this user. This is facilitated by an intermediate step where a search query term is used to map a user to a profile.

Later on, one possible application to recommendations is to suggest the URLs in profiles of previous users with similar search query terms. We call this *On-First-Click-Search-Mediated-Recommendations*: they directly bypass a lengthy navigation session by bridging the gap between pure browsers and premeditated searchers. Another possibility is to map search terms to the documents/content on website. This will transcend clicks/usage access data and map URLs to popular search terms.

5.3 Decision-Tree Based Profile Prediction with Zero Clicks from Search Query Terms

The nearest profile prediction model and all collaborative filtering methods make the critical assumption that a sufficient number of clicks have already been made in the current session. When a user just visits a website, it may be hard to classify the new session based on a single click, particularly if this single click is to a rather general area of the website such as the main page. *In fact even before the user makes any clicks inside the website, it is possible to obtain some idea about the user's information need*. For instance, if the REFERRER for a brand new request indicates that the user has just typed in a query on a search engine, *then this query can be used to infer the information need of the user even prior to making any clicks whatsoever*. Assuming that we have previously trained a classifier to infer the profile number given a search phrase that consists of the terms

appearing in the queries of sessions initiated by *pre-meditated searchers*; then we can classify a brand new session initiated by a new pre-meditated searcher. In this paper, we propose using decision trees [21] for the task of inferring the user profile from the user's search query. Once trained, using the decision tree model to classify a new session is very fast, and constitutes the single step of the recommendation process, since the significant URLs in the classified profile *form* the recommendation set, which serves as a shortcut to more specific areas of the website that by-pass an otherwise lengthy browsing session.

A search query is presented as an input binary vector to the decision tree and a profile is predicted as the output. Each search query term in the input vector is considered as an attribute. In learning, first the entire training data set is presented. Here, an attribute value is tested at each decision node with two possible outcomes of the test, each leading down to a different branch. At the bottom of the tree, we find the leaf or class nodes. A class node indicates the profile to be predicted. An example is illustrated in Figure 6. In our preliminary experiments, we found that an accuracy higher than 90% can easily be achieved in mapping search terms to profiles. However, only a few terms were found to be relevant to the task, while all remaining terms were absent from the decision tree. This may result in low coverage, however this may no be a problem, since this method was intended only to be combined with other methods such as the one in Section 5.1. Alternatively, coverage can be improved if much more data is collected, so that more search terms become part of the prediction model. Another way that this can be improved is by exploiting an existing thesaurus (either a general thesaurus available in the public domain such as WordNet [32], or a specialized thesaurus for the special type of information that is delivered on this website). A thesaurus can be used to extract more general and/or synonymous terms to improve coverage or more specific terms to improve the profile prediction based on a search term. The thesaurus can be exploited both at the model construction phase (by mapping terms) or at the usage/deployment phase as explained above.

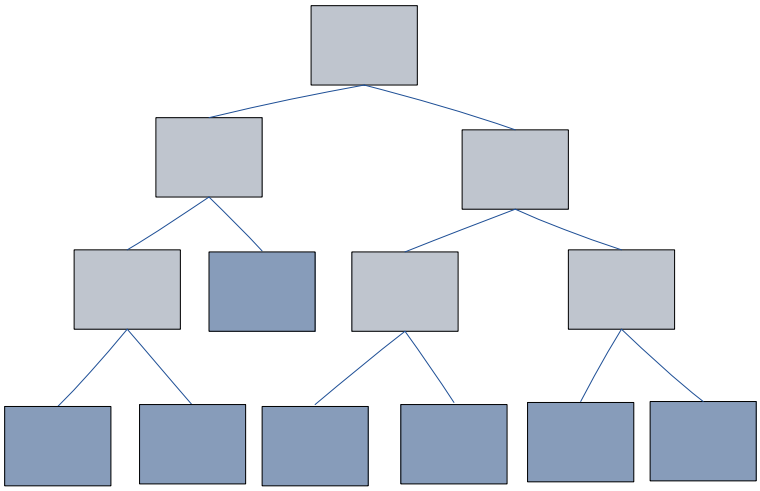

Figure 6: Decision tree example for query term based profile prediction (shaded nodes are the final predictions, lighter nodes are the input attributes/search query terms form the REFERRER field). Branches are labeled with the value of the top node (1 if the term is used in a query, and 0 otherwise)

6. CONCLUSIONS

Understanding the different modes of usage on a heavily visited website can be accomplished using a variety of Web usage mining techniques that can automatically extract frequent access patterns from the user clickstreams stored in web log files. We presented a case study summarizing our preliminary approach and findings in mining web usage patterns from the Web log files of a real life website that has all the challenging aspects of real life web usage mining, including evolving user profiles and access patterns, dynamic web pages, and external data describing the ontology of the web content. Our preliminary results have shown a proof of concept of what can be accomplished with data mining of the web access logs, and how to proceed with data that is added in batches. We have opted to analyze the web logs in batches of two weeks to one month worth of Web log data. Yet, there is no reason why this process could not use either a finer (such as weekly) or coarser (such as quarterly) time resolution, or even an on-demand update of the analysis at any instant in time. We have also opted to analyze non-overlapping batches of access data. However, this does not preclude one from analyzing overlapping batches that

can allow a more continuous dichotomization of the temporality of the web access patterns. We also note that because of space limitations, we do not discuss other pertinent details such as cleaning the Web logs from the massive requests that originate from the various Web crawlers or bots. We also did not discuss our additional efforts in incorporating the content of the web pages into the web usage mining and in enriching the user profiles, as well as the automatic discovery of affiliation information from the users' IP addresses to enrich the personalization process, but plan to do so in the future. Finally, we mention that in this study, we did not focus on the scalability issues, but we plan to address scalability in the future by following an approach similar to the one that we proposed in [22], where the Web clickstreams are considered as a special case of an *evolving data stream*, and where the user profiles are mined in a single pass and in a continuous fashion.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation CAREER Award IIS-0133948. Partial support was also provided by a grant from Innovative Productivity Incorporated.

Table 7: Illustrating the Tracking of Evolving User profiles with A Few Example Profiles (Underlined and bold items indicate information inferred from the website ontology of what would otherwise appear as encoded dynamic URLs. Profiles in the same row are found to be compatible based on their similarity compared to the profile boundaries. Similarity is given by (3), and takes into account the website hierarchical structure and/or relations between different dynamic URLs based on the website ontology). *Birth* of a new profile can be seen when the cells preceding it on the same row are empty (e.g. rows 11 and 12), while *Death* is when the cells following it are empty (e.g. rows 2 and 9). *Atavism* is when two profiles on the same row are separated by some empty cells (e.g. row 5). *Persistent* profiles show activity in contiguous cells on the same row. In this case notice how some profiles can be split into more specific profiles in subsequent periods, and vice versa, some profiles seem to merge into one more general profile with time (e.g. rows 2 and 3)

Row	June 2004	July 2004	Aug 2004-Part 1	Aug 2004-Part 2	Sep 2004
1	Jun4 /Hsarg/Appendixc/C6c2_Sigmacoatingssafetydatasheet.Pdf	Jul4 /Hsarg/Appendixc/C6c2_Sigmacoatingssafetydatasheet.Pdf /Hsarg/Appendixc/C4c_Industrialandmarinecoatings.Pdf	Aug4-Part1 /Hsarg/Appendixc/C6c2_Sigmacoatingssafetydatasheet.Pdf /Hsarg/Appendixc/C3c_Intlmaterialsafetydatasheet.Pdf	Aug4-Part2 /Hsarg/Appendixc/Appc_Contents.Aspx /Hsarg/Appendixc/C6c2_Sigmacoatingssafetydatasheet.Pdf /Hsarg/Appendixc/C5c_Industrialandmarinecoatings.Pdf	Sep4 /Hsarg/Appendixc/Appc_Contents.Aspx /Hsarg/Appendixc/C6c2_Sigmacoatingssafetydatasheet.Pdf
2	Jun4 /Surface.Js /Left1.Htm /Menu.Aspx/Menu_File=Nst.Js /Top_Frame.Aspx /Menu.Aspx/Menu_File=Surface.Js /Mainpage_Frameset.Aspx/Menu_File=Surface.Js&Main=Universal.Aspx **/Surface Treatment/Surface Preparation/Abrasive Blasting** /Frames.Aspx/Menu_File=.Js&	Jul4 /Left1.Htm /Top_Frame.Aspx **/Company Connection/Manufacturers / Suppliers/Soda Blasting** **/Surface Treatment/Surface Preparation/Abrasive Blasting** /Frames.Aspx/Menu_File=.Js&Main=/Universal.Aspx /Mainpage_Frameset.Aspx/Menu_File=.Js&Main=/Universal.Aspx	Aug4-Part1 /Left1.Htm /Top_Frame.Aspx **/Surface Treatment/Surface Preparation/Abrasive Blasting** /Frames.Aspx/Menu_File=.Js&Main=/Universal.Aspx /Mainpage_Frameset.Aspx/Menu_File=.Js&Main=/Universal.Aspx /Menu.Aspx/Menu_File=.Js	Aug4-Part2 /Left1.Htm /Top_Frame.Aspx **/Surface Treatment/Surface Preparation/Abrasive Blasting** /Frames.Aspx/Menu_File=.Js&Main=/Universal.Aspx /Mainpage_Frameset.Aspx/Menu_File=.Js&Main=/Universal.Aspx /Menu.Aspx/Menu_File=.Js	
3	Jun4 /Left1.Htm /Top_Frame.Aspx **/Company Connection/Manufacturers / Suppliers/Soda Blasting** /Frames.Aspx/Menu_File=.Js&Main=/Universal.Aspx /Mainpage_Frameset.Aspx/Menu_File=.Js&Main=/Universal.Aspx /Menu.Aspx/Menu_File=.Js	Jul4 /Left1.Htm /Top_Frame.Aspx **/Company Connection/Manufacturers / Suppliers/Soda Blasting** **/Surface Treatment/Surface Preparation/Abrasive Blasting** /Frames.Aspx/Menu_File=.Js&Main=/Universal.Aspx /Mainpage_Frameset.Aspx/Menu_File=.Js&Main=/Universal.Aspx	Aug4-Part1 /Left1.Htm /Top_Frame.Aspx **/Surface Treatment/Surface Preparation/Abrasive Blasting** /Frames.Aspx/Menu_File=.Js&Main=/Universal.Aspx /Mainpage_Frameset.Aspx/Menu_File=.Js&Main=/Universal.Aspx /Menu.Aspx/Menu_File=.Js	Sep4 /Left1.Htm /Top_Frame.Aspx **/Surface Treatment/Surface Preparation/Abrasive Blasting** /Frames.Aspx/Menu_File=.Js&Main=/Universal.Aspx /Mainpage_Frameset.Aspx/Menu_File=.Js&Main=/Universal.Aspx /Menu.Aspx/Menu_File=.Js	
4		Jul4 /Left1.Htm /Top_Frame.Aspx **/Surface Treatment/Standards And Specifications/Navsea Approved Ppis** **/Navy Community/Navy/Technical Needs & Commercial Solutions/U.S. Navy Underwater Hull Anti Fouling Coating** **/Surface Treatment/Surface Preparation/Abrasive**	Aug4-Part1 /Left1.Htm /Top_Frame.Aspx **/Surface Treatment/Surface Preparation/Abrasive Blasting** /Frames.Aspx/Menu_File=.Js&Main=/Universal.Aspx /Mainpage_Frameset.Aspx/Menu_File=.Js&Main=/Universal.Aspx /Menu.Aspx/Menu_File=.Js		

5	Jun4 /Userwebhelp/Whgdata/Whnvl32.Htm	Jul4 /Userwebhelp/Whgdata/Whnvt30.Htm /Userwebhelp/Whgdata/Whlstt0.Htm /Userwebhelp/Whgdata/Whnvp30.Htm /Userwebhelp/Whgdata/Whlsti0.Htm /Userwebhelp/Whgdata/Whlstf0.Htm			Sep4 /Userwebhelp/Whgdata/Whlstf4.Htm /Userwebhelp/Whgdata/Whlstf5.Htm /Userwebhelp/Whgdata/Whlstt5.Htm /Userwebhelp/Whgdata/Whlstfl13.Htm /Userwebhelp/Offsite_Links.Htm /Userwebhelp/Whgdata/Whnvf31.Htm /Userwebhelp/Whgdata/Whnvt30.Htm /Userwebhelp/Whgdat
6	Jun4 /Userwebhelp/Nstc_Webhelp.Htm /Userwebhelp/Using_Nstm.Htm /Userwebhelp/Whmsg.Js /Userwebhelp/Whproxy.Js /Userwebhelp/Whver.Js /Userwebhelp/Whutils.Js /Userwebhelp/Whtopic.Js /Userwebhelp/Whstub.Js /Userwebhelp/Whstart.Js /Userwebhelp/Whskin_Tbars.Htm	Jul4 /Userwebhelp/Nstc_Webhelp.Htm /Userwebhelp/Using_Nstm.Htm /Userwebhelp/Whmsg.Js /Userwebhelp/Whproxy.Js /Userwebhelp/Whver.Js /Userwebhelp/Whutils.Js /Userwebhelp/Whtopic.Js /Userwebhelp/Whstub.Js /Userwebhelp/Whstart.Js /Userwebhelp/Whskin_Tbars.Htm	Aug4-Part1 /Userwebhelp/Nstc_Webhelp.Htm /Userwebhelp/Using_Nstm.Htm /Userwebhelp/Whmsg.Js /Userwebhelp/Whproxy.Js /Userwebhelp/Whver.Js /Userwebhelp/Whutils.Js /Userwebhelp/Whtopic.Js /Userwebhelp/Whstub.Js /Userwebhelp/Whstart.Js /Userwebhelp/Whskin_Tbars.Htm	Aug4-Part2 /Userwebhelp/Nstc_Webhelp.Htm /Userwebhelp/Using_Nstm.Htm /Userwebhelp/Whmsg.Js /Userwebhelp/Whproxy.Js /Userwebhelp/Whver.Js /Userwebhelp/Whutils.Js /Userwebhelp/Whtopic.Js /Userwebhelp/Whstub.Js /Userwebhelp/Whstart.Js /Userwebhelp/Whskin_Tbars.Htm	Sep4 /Userwebhelp/Nstc_Webhelp.Htm /Userwebhelp/Using_Nstm.Htm /Userwebhelp/Whmsg.Js /Userwebhelp/Whproxy.Js /Userwebhelp/Whver.Js /Userwebhelp/Whutils.Js /Userwebhelp/Whtopic.Js /Userwebhelp/Whstub.Js /Userwebhelp/Whstart.Js /Userwebhelp/Whskin_Tbars.Htm
7	Jun4 /Rsandp/Presentations/Fleet+Corrosion+Control+Forum+" /Rsandp/Presentations/Advances+In+Technology+And+Standards+For+Mitigating+The+Effect+Of+Soluble+Salt.Pdf /Rsandp/Presentations/The+Role+Of+Relative+Humidity+In+Corrosion.Pdf	Jul4 /Rsandp/Presentations/The+Role+Of+Relative+Humidity+In+Corrosion.Pdf /Rsandp/Presentations/Review+Of+Conductive+Polymers+As+Benign+Corrosion+Control+Coatings.Pdf	Aug4-Part1 /Rsandp/Presentations/Whats+New+In+The+Fight+Against+Cavitation.Pdf /Rsandp/Presentations/Advances+In+Technology+And+Standards+For+Mitigating+The+Effect+Of+Soluble+Salt.Pdf /Rsandp/Presentations	Aug4-Part2 /Rsandp/Presentations/Fleet+Corrosion+Control+Forum+ /Rsandp/Presentations/Future+Naval+Capability+	Sep4 /Rsandp/Presentations/Advances+In+Technology+And+Standards+For+Mitigating+The+Effect+Of+Soluble+Salt.Pdf /Rsandp/Presentations /Rsandp/Presentations/An+Evaluation+Of+The+Corrosion+Resistance+Of+Thermal
8		Jul4 /Left1.Htm /Top_Frame.Aspx **/Write_Up_Display.Aspx/Navy Community/Navy/Approved Interior Coatings&Write_Ups_Id/Bilge Preservation** **/Write_Up_Display.Aspx/Navy Community/Navy/Approved Interior Coatings&Write_Ups_Id/Ballast Tanks** /Menu.Aspx/Menu_File=.Js		Aug4-Part2 **/Write_Up_Display.Aspx/Navy Community/Navy/Approved Interior Coatings&Write_Ups_Id/Bilge Preservation** **/Write_Up_Display.Aspx/Write_Ups_Id/Project: MSC Non-Skid** **/Write_Up_Display.Aspx/Navy Community/Navy/Approval Of Coatings For Navy Use (Roadmap)&Wr**i	Sep4 /Left1.Htm /Top_Frame.Aspx **/Write_Up_Display.Aspx/Navy Community/Navy/Approved Interior Coatings&Write_Ups_Id/Bilge Preservation** **/Write_Up_Display.Aspx/Navy Community/Navy/Approved Exterior Coatings&Write_Ups_Id/Non-Skid Walk** **/Menu.Aspx/Menu_File=.Js**
9		Jul4 /Event_Details.Aspx/Id=**Control Systems 2004 Conference** /Event_Details.Aspx/Id=**2004 Gordon Conference On Aqueous Corrosion** /Event_Details.Aspx/Id=**SSPC NAVSEA Basic Paint Inspector (NBPI) Course** /Event_Details.Aspx/Id=**Rust 2004 14th Corrosion Technolog**			Sep4 /Event_Details.Aspx/Id=**The 27th Annual Workshop On Electrochemical Measurements**
10		Jul4 /Write_Up_Display.Aspx/Write_Ups_Id/**Project: MSC Non-Skid** /Write_Up_Display.Aspx/Write_Ups_Id/**Project: Navy Rudder Coating Failures** /Write_Up_Display.Aspx/Write_Ups_Id/**Project: UV Holiday Detection** /Write_Up_Display.Aspx/**Navy Community/Navy/Approved**	Aug4-Part1 /Write_Up_Display.Aspx/Write_Ups_Id/**Project: UV Holiday Detection**		
11					Sep4 **/Usaf/Gallery**/Pages/Dsc01010_Jpg.Htm /Usaf/Gallery/Pages/Dsc01009_Jpg.Htm /Usaf/Gallery/Pages/Dsc01008_Jpg.Htm /Usaf/Gallery/Pages/Dsc01007_Jpg.Htm /Usaf/Gallery/Pages/Dsc01006_Jpg.Htm /Usaf/Gallery/Pages/Dsc01013_Jpg.Htm /Usaf/Gallery/Pages/Dsc01012

12					Sep4 /Default.Htm /Favicon.Ico /Nst.Js /Navy.Js /Mainpage_Frameset.Aspx/Menu_File=Nst.Js&Main=Nst.Aspx /Left1.Htm /Menu.Aspx/Menu_File=Nst.Js /Top_Frame.Aspx /Nst.Aspx **/Nst Center&Reg;/Regulations And Laws** /Menu.Aspx/Menu_File=Navy.Js /Mainpage_Frameset.A

REFERENCES

[1] M. Perkowitz and O. Etzioni. Adaptive web sites: Automatically learning for user access pattern. Proc. 6th int. WWW conference, 1997.

[2] R. Cooley, B. Mobasher, and J. Srivastava. Web Mining: Information and Pattern discovery on the World Wide Web, Proc. IEEE Intl. Conf. Tools with AI, Newport Beach, CA, pp. 558-567, 1997.

[3] O. Nasraoui and R. Krishnapuram, and A. Joshi. Mining Web Access Logs Using a Relational Clustering Algorithm Based on a Robust Estimator, 8th International World Wide Web Conference, Toronto, pp. 40-41, 1999.

[4] O. Nasraoui, R. Krishnapuram, H. Frigui, and A. Joshi. Extracting Web User Profiles Using Relational Competitive Fuzzy Clustering, International Journal on Artificial Intelligence Tools, Vol. 9, No. 4, pp. 509-526, 2000.

[5] O. Nasraoui, and R. Krishnapuram. A Novel Approach to Unsupervised Robust Clustering using Genetic Niching, Proc. of the 9th IEEE International Conf. on Fuzzy Systems, San Antonio, TX, May 2000, pp. 170-175.

[6] O. Nasraoui and R. Krishnapuram. A New Evolutionary Approach to Web Usage and Context Sensitive Associations Mining, International Journal on Computational Intelligence and Applications - Special Issue on Internet Intelligent Systems, Vol. 2, No. 3, pp. 339-348, Sep. 2002.

[7] M. Pazzani and D. Billsus. Learning and revising User Profiles: The identification of Interesting Web Sites, Machine Learning, Arlington, 27.

[9] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective personalizaton based on association rule discovery from Web usage data, ACM Workshop on Web information and data management, Atlanta, GA, Nov. 2001.

[10] J. H. Holland. *Adaptation in natural and artificial systems*. MIT Press, 1975.

[13] R. Agrawal and R. Srikant. Fast algorithms for mining association rules, Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994, pp. 487-499.

[14] G. Linden, B. Smith, and J. York. *Amazon.com* Recommendations Item-to-item collaborative filtering, IEEE Internet Computing, Vo. 7, No. 1, pp. 76-8

[15] J. Breese, H. Heckerman, and C. Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering, Proc. 14th Conf. Uncertainty in Artificial Intelligence, pp. 43-52, 1998.

[16] J.B. Schafer, J. Konstan, and J. Reidel. Recommender Systems in E-Commerce, Proc. ACM Conf. E-commerce, pp. 158-166, 1999.

[17] J. Srivastava, R. Cooley, M. Deshpande. and P-N Tan, Web usage mining: Discovery and applications of usage patterns from web data, SIGKDD Explorations, Vol. 1, No. 2, Jan 2000, pp. 1-12.

[18] O. Zaiane, M. Xin, and J. Han. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs, in "Advances in Digital Libraries", 1998, Santa Barbara, CA, pp. 19-29.

[19] M. Spiliopoulou and L. C. Faulstich. WUM: A Web utilization Miner, in Proceedings of EDBT workshop WebDB98, Valencia, Spain, 1999.

[20] J. Borges and M. Levene, Data Mining of User Navigation Patterns, in *"Web Usage Analysis and User Profiling"*, Lecture Notes in Computer Science", H. A. Abbass, R. A. Sarker, and C.S. Newton Eds., Springer-Verlag,1999 , pp. 92-111.

[21] J. R. Quinlan. Induction of Decision Trees. Machine Learning, Vol. 1, pp. 81--106, 1986.

[22] O. Nasraoui, C. Cardona, C. Rojas, and F. Gonzalez. Mining Evolving User Profiles in Noisy Web Clickstream Data with a Scalable Immune System Clustering Algorithm, in Proc. of WebKDD 2003, Washington DC, Aug. 2003, 71-81.

[23] G. Adomavicius, A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. Knowl. Data Eng. 17(6): 734-749, 2005.

[24] M. Pazzani. A Framework for Collaborative, Content-Based and Demographic Filtering, AI Review, 1999.

[25] M. Balabanovic and Y. Shoham. Fab: Content-based, Collaborative Recommendation, Communications of the ACM 40(3), March 1997.

[26] B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining. In Proc. International Semantic Web Conference (ISWC02), 2002.

[27] R. Burke. Hybrid recommmender systems: Survey and experiments. In User Modeling and User-Adapted Interaction, 2002.

[28] D. Oberle, B. Berendt, A. Hotho, and J. Gonzalez. Conceptual User Tracking, in Proc. of the Atlantic Web Intelligence Conference (AWIC) Madrid, Spain, 2003.

[29] H. Dai and B. Mobasher. Using ontologies to discover domain-level web usage profiles. In Proc. 2nd Semantic Web Mining Workshop at ECML/PKDD-2002.

[30] M. Eirinaki, H. Lampos, M. Vazirgiannis, I. Varlamis. SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process, in the Proc. of SIGKDD '03, Washington DC, USA, August 2003.

[31] P. Van der Putten, J. N. Kok and A. Gupta. Why the Information Explosion Can Be Bad for Data Mining and How Data Fusion Provides a Way Out, In Proc. of the 2nd SIAM International Conference on Data Mining, 2002.

[32] Miller, G. A. WORDNET: An On-Line Lexical Database, Int. Journal of Lexicography 3-4:235-312, 1990.