Concept-based Clustering of Clickstream Data

Arindam Banerjee and Joydeep Ghosh
{abanerje,ghosh}@ece.utexas.edu
Dept. of Electrical and Computer Engineering
University of Texas at Austin
Austin, TX 78712

Abstract— **Determining the *type* of a user based on his interaction with a website is a key problem in web usage mining. Absence of proper cookie information makes the task more difficult since one then has to extract exact session information for the anonymous users before performing any advanced analytics such as user characterization. In this paper, we first discuss two ways of session extraction from weblog data. Then we propose a novel and effective algorithm for clustering webusers that takes into account both the trajectory taken through a website and the time spent at each page. Results are presented on weblogs of www.sulekha.com to illustrate the techniques.**

keywords : web usage mining, clickstream, session extraction, clustering, data mining.

I. Introduction

With the rapid increase in web-traffic and e-commerce, understanding user behavior based on their interaction with a website is becoming more and more important for website owners. Identifying the nature of each user surfing a website may enable that site to provide customized content for the users, thereby making it more *sticky* and enhancing user experience. The business implications of such an ability are huge, specially for portals, personalized content providers and e-tailers. Several techniques have been proposed for this problem [1], [2], [3], [4] but a definitive solution is yet to emerge.

The footprint that a webuser generates at a particular website is his "*cowpath*" in that site. Identifying the category of a user from his cowpath is a very difficult problem since the cowpath is across multiple webpages. Moreover the time spent at each page in his path varies. We have developed a formal definition of a cowpath and a way to compute similarity values between any two cowpaths for a given resolution taking into account both the trajectory taken and the time spent at each page.

If a user is not tracked by user and session cookies, session extraction is necessary before any further processing such as cowpath analysis of the log file can be done. This procedure has become even more important because of privacy concerns that are expected to limit future use of cookie information. In this paper we suggest two techniques for session extraction.

Clustering of the extracted sessions is performed in a suitable similarity space using efficient graph partitioning techniques. Consider a path feature space to be the set of all possible paths. Then each session covers a subset of this set. While computing the similarity between any two sessions, we first determine the intersection of the two subsets corresponding to the two sessions that are being compared and then perform the similarity computations primarily in the intersection set using time information from each session. If the subsets do not intersect, the similarity value is zero. A graph is constructed whose nodes are the paths and the edge connecting any two nodes is the similarity value between the two paths – this forms the similarity space. Finally we use graph partitioning to get the clusters of paths.

For large websites, if paths are considered at a webpage level of resolution, many similarity values are zero since very few paths have actual page overlaps. This does not reflect session behavior properly. Therefore we introduce the idea of concept-based webpaths so that sessions that surfed through conceptually similar pages have a non-null intersection even without any direct webpage overlap. This procedure is effective in yielding meaningful clusters of user sessions.

The paper is organized as follows. Section II discusses the preprocessing that is done on the logfiles. In section III we discuss two techniques for extracting sessions of anonymous users. In section IV we formally define a cowpath and suggest a way of computing similarity between two cowpaths. In section V we extend the idea of a cowpath to a concept based cowpath. In section VI we study the flexibility and scalability of the algorithm. Section VII discusses some sample results. We draw some conclusions in section VIII.

II. Preprocessing

Preprocessing is a very important step in any kind of log analysis. Normally one is interested in processing page-views rather than page-hits. When a user accesses

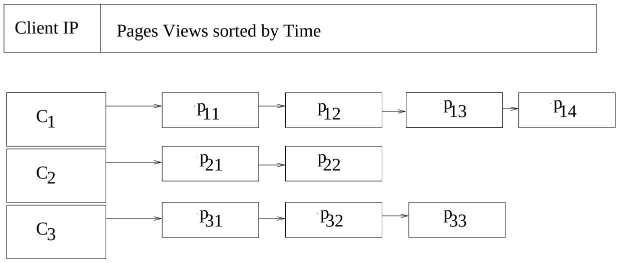


Fig. 1. The hash representation of the weblog

a webpage, typically multiple files such as associated images (jpg,gif files) are accessed from the server. So a particular page-view is actually recorded as multiple lines of webserver log. Moreover, there are access error entries in the log file which are not of interest. Also, the weblog entries corresponding to dynamically generated webpages from cgi-scripts and those corresponding to *post* operations are difficult to analyze since they do not have any static existence in the website hierarchy. In the preprocessing stage we filter out the image file entries, the access error entries and the entries generated from cgi-binaries and *post* operations. We also filter out any entries corresponding to known webcrawlers and local bots helping the onsite search engine. The trimmed weblog file is normally anywhere from $\frac{1}{4}$th to $\frac{1}{10}$th of the original raw weblog file, in terms of the number of entries.

III. Session Extraction

For user session extraction, the weblog information is stored in a hash-table with the IP address of the client as the hash key. In addition, the collision list corresponding to each accessing client is stored as a time sorted list of pages accessed from that client. Fig. 1 shows the hash representation of the weblog. One should note that there may be multiple users accessing the same website from the same client at different times. Moreover, each user may have multiple content access windows open for the same site at a particular time or maybe using different clients to access the same site. So it is very difficult to separate the cowpaths of the different users from the same client. For anonymous users, one is interested in extraction of various sessions from a particular client since there is no way to ascribe sessions to actual users. We propose two heuristics for doing this.

A. Time-out heuristic

The basic idea in this approach is that if there are no hits from a particular client for a time τ, then session is assumed to have ended. Any hit after the time-out will be ascribed to a new session. The new session will continue until the dormancy period between successive hits from that client is more than the time-out threshold τ. The assumption in this approach is that there is only user at a time from a client, which may not always be valid. We used routines from [5] for this stage.

B. Link following heuristic

This approach is more involved in the sense that it tries to extract the exact path of a user from a client making use of the link structure of the website. A simple depth-first webcrawler starting from the home page of the website is used to generate a directed graph of the site where the nodes are the webpages in the site and there is a edge from node n_1 to n_2 iff there is a hyperlink from n_1 to n_2. The webcrawler algorithm works as follows :

Let $\mathbf{L}$ be a stack of webpages. Let $\mathbf{H}$ be a hash such that the keys of $\mathbf{H}$ are the unique webpages in the domain and the satellite-data is 1 or 0 depending on whether that page has been visited by the crawler or not. For any webpage ω, we define the function $\textsc{Getpage}(\omega)$ such that $R_\omega \leftarrow \textsc{Getpage}(\omega)$, where R_ω is the set of hyperlinks in the page belonging to the same website. We also define the function $\textsc{Directed-Edge}(\omega, R_\omega)$, which generates directed edges from ω to the webpages corresponding to each element in R_ω. If ω_0 be the starting page, normally the root homepage, e.g., http://www.foo.com/index.html, then the algorithm followed by the crawler is as follows :

```
$\textsc{Push}(\mathbf{L}, \omega_0)$
$\textsc{Keys}(\mathbf{H}) \leftarrow \omega_0$
$\textit{while}\ (\textsc{Not}\ (\textsc{Stack-Empty}(\mathbf{L})))$
    $\omega = \textsc{Pop}(\mathbf{L})$
    $R_\omega \leftarrow \textsc{Getpage}(\omega)$
    $\textsc{Directed-Edge}(\omega, R_\omega)$
    $\forall \omega' \in R_\omega, \textit{if}\ \omega' \notin \textsc{Keys}(\mathbf{H})$
            $\textsc{Push}(\mathbf{L}, \omega')$
            $\mathbf{H}\{\omega'\} \leftarrow 1$
    $\textit{end}$
$\textit{end}$
```

This creates a directed graph corresponding to the website.

To extract a session from a particular client, we check whether there are links in the website graph corresponding to successive hits from the client. For example, from Fig. 1, for the client machine C_1, we check for validity of links between pages p_{11} and p_{12}. If p_{11} has a edge to p_{12} in the website graph, we consider that as a valid transition, put them in the same session and check for validity of transition from p_{12} to p_{13}. If a there is no edge connecting p_{11} and p_{12}, we mark page p_{12} as not visited in this session and go on to the check for the validity between p_{11} and p_{13}. At the end of the first pass we get one full session from C_1; then we take all the pages that are marked as not visited in the first session and continue the process till the list is exhausted. The assumption in this approach is that the users never types out a webpage in full or press the browser provided buttons like back, forward etc., which is not always true.

The time-out heuristic is much faster than the link following heuristic since it does not have to crawl the website and does not have to do website graph adjacency matrix lookups[1]. However, the link following helps in extracting exact sessions as long as the user does not press browser provided buttons or type in urls. We found that the time-out heuristic was more effective perhaps because the assumptions it makes are more realistic.

IV. Cowpath Similarity Measure

A very important aspect of clickstream analysis is clustering of users based on their cowpaths. The purpose of clustering users based on the paths they have traced out in a particular website is to find groups of users with similar interests and motivations for visiting the website. If the site is well designed there will be strong correlation between the similarity of user navigation paths and similarity among the users' interests or intentions. Therefore, clustering of the former could be used to predict clusters for the latter.

Before defining the similarity between cowpaths, we note that a cowpath of a user consists of not only a sequence of webpages, but also the time spent in each webpage. The time spent by a user in a particular page is a clear indication of his interest in the page[2].

The paths generated in the sessions are of varying lengths and can start on almost any page in the website. So, in order to measure the similarity between paths, we first conceptualize a path feature space which is a set consisting of all possible paths for that website. At first the subset of this feature space covered by any given path will be defined. Then, comparing any two paths essentially reduces to finding the similarity between the paths in the intersection set of the subsets covered by each of the paths, normalized by the size of the subsets.

A. Path Feature Space

Consider an N-hop cowpath $\alpha = [(\alpha_0, \tau_{\alpha_0}) \ \ (\alpha_1, \tau_{\alpha_1}) \cdots (\alpha_N, \tau_{\alpha_N})]$ traced out in a session, where α_i is the ith page visited in this session and τ_{α_i} is the time spent on that page. Let $S_\alpha^{(\mu)}$ be the set of all μ-hop subpaths of α:

$$S_\alpha^{(\mu)} = \{(A_0^{(\mu)}, T_{A_0^{(\mu)}}), (A_1^{(\mu)}, T_{A_1^{(\mu)}}) \cdots (A_{N-\mu}^{(\mu)}, T_{A_{N-\mu}^{(\mu)}})\} \tag{1}$$

where $\mu = 0, 1, \cdots, N$ and for $n = 0 \cdots (N-\mu)$,

$$A_n^{(\mu)} = A_n^{(\mu)}(\alpha) = \alpha_n \alpha_{n+1} \cdots \alpha_{n+\mu} \tag{2}$$

$$T_{A_n^{(\mu)}} = T_{A_n^{(\mu)}}(\alpha) = \sum_{i=n}^{n+\mu} \tau_{\alpha_i} \tag{3}$$

So, $A_n^{(\mu)}$ is the sequence of $(\mu+1)$ successive pages in the session starting from α_n, the nth page in the session, and $T_{A_n^{(\mu)}}$ is the total time spent in this subpath of length $(\mu+1)$. The union of all the sets $S_\alpha^{(\mu)}$ for $\mu = 0, 1, \ldots, N$, forms the feature subset covered by the cowpath α and is given by :

$$\mathcal{F}(\alpha) = \bigcup_{\mu=0}^{N} S_\alpha^{(\mu)} \tag{4}$$

We note that the cardinality of $S_\alpha^{(\mu)}$ decreases linearly with an increase in μ. Since we are interested in the similarity between paths, we never actually compute the feature subspace spanned by a path. Instead, for any pair of paths, we compute their intersection space at the μ-hop level, $\forall\mu$, which is typically much smaller than the feature subspace for either of the cowpaths.

B. Min-Max Path Similarity

The similarity measure used here takes into account the time spent as in [3] but with a very important refinement. In [3] the product of the times spent as a similarity measure. This returns the same similarity values for $t_1 = 1, t_2 = 100$ and $t_1 = 10, t_2 = 10$. Our measure gets rid of such problems and adds a lot of flexibility to the formulation.

Given any two cowpaths α and β, we first define their intersection space at the μ-hop level as

$$\begin{aligned} \Psi_{\alpha,\beta}^{(\mu)} &= S_\alpha^{(\mu)} \wedge S_\beta^{(\mu)} \\ &= \{\psi_i^{(\mu)} \mid \exists n, m \text{ such that} \\ &\quad\ \psi_i^{(\mu)} = A_n^{(\mu)}, (A_n^{(\mu)}, T_{A_n^{(\mu)}}) \in S_\alpha^{(\mu)}, \\ &\quad\ \psi_i^{(\mu)} = B_m^{(\mu)}, (B_m^{(\mu)}, T_{B_m^{(\mu)}}) \in S_\beta^{(\mu)}\} \end{aligned} \tag{5}$$

where $\wedge$ stands for the intersection between the first entry of the sets having two-tuple elements. The similarity between cowpaths α and β at the μ-hop level is then given by :

$$\rho_{\alpha,\beta}^{(\mu)} = \sum_{\substack{\psi_i^{(\mu)} \in \Psi_{\alpha,\beta}^{(\mu)} \\ A_n^{(\mu)}, B_m^{(\mu)} = \psi_i^{(\mu)}}} \frac{\min(T_{A_n^{(\mu)}}, T_{B_m^{(\mu)}})}{\max(T_{A_n^{(\mu)}}, T_{B_m^{(\mu)}})} \cdot \frac{T_{A_n^{(\mu)}}}{T_\alpha} \cdot \frac{T_{B_m^{(\mu)}}}{T_\beta} \tag{6}$$

where the min-max component computes the time-similarity on the matched subpath $\psi_i^{(\mu)}$ and the other two time-fraction components weigh this similarity by the

[1] this operation is O(1), after the graph has been generated

[2] however one needs to set a heuristic upper bound on the time spent on any page by any user to take care of outliers, distracted users, etc.

importance each of the sessions attach to this subpath similarity in terms of the fraction of the total session time spent in this subpath. The overall similarity between α and β is given by

$$< \alpha, \beta >= \sum_{\mu=0}^{\min(N,M)} w_\mu \rho^{(\mu)}_{\alpha,\beta} \tag{7}$$

where w_μ is the normalized weight of the similarity at the μ-hop level. For our algorithm, the normalized weight vector $\vec{w} = [w_0, w_1, \cdots, w_\mu, \cdots, w_{\min(N,M)}]$ is a user defined parameter. It gives the user the flexibility to look at the path-similarities at varying resolutions. Finally, the normalized similarity between the cowpaths α and β is given by:

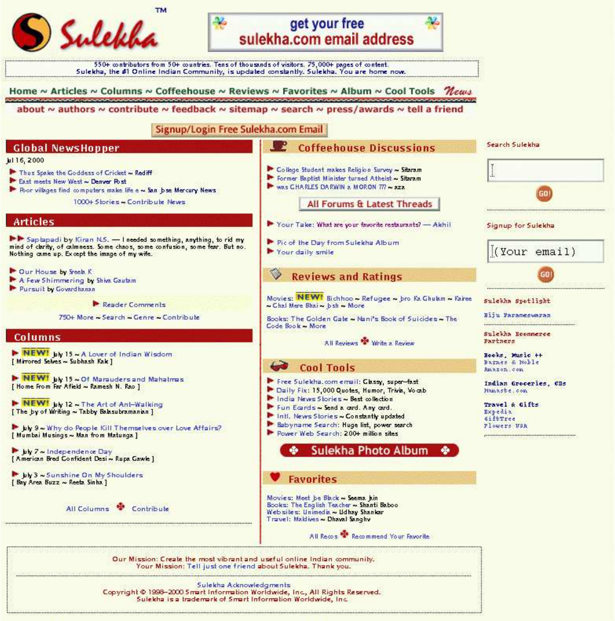
$$\cos(\theta(\alpha, \beta)) = \frac{< \alpha, \beta >}{(< \alpha, \alpha >)^{\frac{1}{2}} (< \beta, \beta >)^{\frac{1}{2}}} \tag{8}$$

V. Concept-based Clustering

The results obtained from grouping cowpaths at page resolution are often difficult to interpret as such a fine resolution leads to a wide variety of cowpaths, many of which are semantically "equivalent". To tackle this problem, the webpages can be first grouped into categories based on suitable metadata information, and then the cowpaths can be formed from the concept-category of the pages present in the cowpath. In particular, since several websites such as *Yahoo* and *Excite* have their contents already categorized based on subject or topic, this structure can be readily used to form the concepts. We have used this method successfully on the weblogs from www.sulekha.com, a highly trafficked community site and from www.ece.utexas.edu, the website of the department of ECE at UT, Austin. The concept-categories for *sulekha* will be discussed next, followed by a few examples of conversion from webpage based cowpaths to concept-based cowpaths.

Fig 2 shows the homepage of *Sulekha*. Being a very structured website, we identified the concept-categories to be the first level branchings from the root page, the root page in itself being a category. We identified the following categories – home, which stands for www.sulekha.com/index.html; articles, authors, biztech, books, coffeehouse, contests, cooltools, creative, fun, games, movies, personal, philosophy, politics, sports, wo-men, each of which corresponds to a first level branching from home, e.g., the pages www.sulekha.com/books/foo.html, www.sulekha.com/books/foo/bar.html fall into the category books; and misc, which stands for all the webpages in the site that do not fall in any of the previously mentioned categories.

When we convert the raw cowpaths to more concept-cowpaths, the average size of a cowpath is reduced and we get cowpaths which can be easily understood.



Sulekha™

get your free
sulekha.com email address

550+ contributors from 50+ countries. Tens of thousands of visitors. 75,000+ pages of content.
Sulekha, the #1 Online Indian Community, is updated constantly. Sulekha. You are home now.

Home ~ Articles ~ Columns ~ Coffeehouse ~ Reviews ~ Favorites ~ Album ~ Cool Tools News

about ~ authors ~ contribute ~ feedback ~ sitemap ~ search ~ press/awards ~ tell a friend

Signup/Login Free Sulekha.com Email

Global NewsHopper

Jul 16, 2000

- Thus Spake the Goddess of Cricket ~ Rediff
- East meets New West ~ Denver Post
- Poor villages find computers make life e ~ San Jose Mercury News

1000+ Stories ~ Contribute News

Articles

Saptapadi by Kiran N.S. — I needed something, anything, to rid my mind of clarity, of calmness. Some chaos, some confusion, some fear. But no. Nothing came up. Except the image of my wife.

- Our House by Sreela K
- A Few Shimmering by Shiva Gautam
- Pursuit by Govardhanan

Reader Comments

750+ More ~ Search ~ Genre ~ Contribute

Columns

NEW! July 15 ~ A Lover of Indian Wisdom
[Mirrored Selves ~ Subhash Kak]

NEW! July 15 ~ Of Marauders and Mahatmas
[Home From Far Afield ~ Ramesh N. Rao]

NEW! July 12 ~ The Art of Ant-Walking
[The Joy of Writing ~ Tabby Balasubramanian]

July 9 ~ Why do People Kill Themselves over Love Affairs?
[Mumbai Musings ~ Man from Matunga]

July 7 ~ Independence Day
[American Bred Confident Desi ~ Rupa Gawle]

July 3 ~ Sunshine On My Shoulders
[Bay Area Buzz ~ Reeta Sinha]

All Columns ✤ Contribute

Coffeehouse Discussions

- College Student makes Religion Survey ~ Sitaram
- Former Baptist Minister turned Atheist ~ Sitaram
- was CHARLES DARWIN a MORON ??? ~ aza

All Forums & Latest Threads

- Your Take: What are your favorite restaurants? — Akhil
- Pic of the Day from Sulekha Album
- Your daily smile

Reviews and Ratings

Movies: NEW! Bichhoo ~ Refugee ~ Joro Ka Ghulam ~ Kairee ~ Chal Mere Bhai ~ Josh ~ More

Books: The Golden Gate ~ Nani's Book of Suicides ~ The Code Book ~ More

All Reviews ✤ Write a Review

Cool Tools

- Free Sulekha.com email: Classy, super-fast
- Daily Fix: 15,000 Quotes, Humor, Trivia, Vocab
- India News Stories ~ Best collection
- Fun Ecards ~ Send a card. Any card.
- Intl. News Stories ~ Constantly updated
- Babyname Search: Huge list, power search
- Power Web Search: 200+ million sites

Sulekha Photo Album

Favorites

Movies: Meet Joe Black ~ Seema Jain
Books: The English Teacher ~ Shanti Baboo
Websites: Unimedia ~ Udhay Shankar
Travel: Maldives ~ Dhaval Sanghv

All Recos ✤ Recommend Your Favorite

Search Sulekha

GO!

Signup for Sulekha

(Your email)

GO!

Sulekha Spotlight

Biju Parameswaran

Sulekha Ecommerce Partners

Books, Music ++
Barnes & Noble
Amazon.com

Indian Groceries, CDs
Namaste.com

Travel & Gifts
Expedia
GiftTree
Flowers USA

Our Mission: Create the most vibrant and useful online Indian community.
Your Mission: Tell just one friend about Sulekha. Thank you.

Sulekha Acknowledgments
Copyright © 1998–2000 Smart Information Worldwide, Inc., All Rights Reserved.
Sulekha is a trademark of Smart Information Worldwide, Inc.

Fig. 2. The home page of *Sulekha*, www.sulekha.com

Once these paths were formed, we computed the similarity between each pair of paths using the min-max path similarity method for a fixed weighting scheme. The similarity matrix is then viewed as an adjacency matrix of a graph where each path is a node and the edge connecting any two nodes is the min-max similarity value between the two paths. Partitioning this graph with min-cut constraints gives k sets of relatively closely connected nodes which are essentially the k clusters from the initial set of nodes. Since the nodes represent the paths, we end up in getting k clusters of the paths. After clustering the paths in this way by using an efficient and fast graph partitioning algorithm called Metis [6], we found that the results were much more meaningful than the simple page-based path case. In section VII, we present the clusters we obtained from concept-based cowpaths using min-max path similarity with linearly decreasing weights for the *sulekha* weblog data. Corresponding to each cluster, we have assigned a meaningful cluster label after looking at the general nature of cowpaths in that cluster. The format of a cowpath in the results is a sequence of <`category,time-spent-in-this-category`> tuples.

VI. A Few Comments

Let us take a closer look at the techniques we have presented so far in terms of the flexibility and the scalability they provide.

✎ **Flexibility** : The weight vector $\vec{w}$ is chosen by the analyst. So, if one is interested in the detail behavior of users in the different pages, the weight vector should be decreasing, thereby giving more weight to subpaths of smaller number of hops. The extreme case of this will be to look at the paths only at a single page resolution, which is a 0-hop and is without any sequence information. In other words, $\vec{w} = [1\ 0\ 0\ \cdots\ 0]$ computes the similarity between two sessions based on just the pages they have visited irrespective of the order and the analysis boils down to a market-basket approach. On the other hand, if one is interested in getting clusters based on more sequence-based information, the weight vector should be increasing, thereby giving more importance to longer subpaths. The extreme case of this will be to look at the sequence of all the pages visited in a session and the total time span of the session. In other words, $\vec{w} = [0\ \cdots 0\ 1]$ computes the similarity between sessions based on exact sequential match of the smaller session with the bigger one. All intermediate values of $\vec{w}$ capture the similarities at various subpath resolutions. For example, $\vec{w} = [\frac{1}{k} \cdots \frac{1}{k}]$, where $k = \min(N, M)$, gives equal importance to similarities between the sessions at all possible subpath resolutions.

- **Scalability** : The path-similarity algorithm essentially computes the similarity between each pair of a total of P paths. The average-case complexity is $O(P^2\bar{L})$, where P is the number of cowpaths, and $\bar{L} = E(YZ^2)$, where $Y = \max(X_1, X_2), Z = \min(X_1, X_2)$, and X_1, X_2 are random variables for the length of a cowpath. In comparison to the raw cowpath with an average length of around 12-15 hops, the concept-clustering approach in which group of pages map to a concept or *metapage*, the average length of a cowpath reduces significantly, resulting in a significant reduction in complexity.

VII. Sample Results

Due to lack of space, we just present the clustering results obtained after doing concept-based clustering on the *sulekha* website. The data used has 184 MB of raw logs, collected over the period Feb 01, 2000 to Feb 29, 2000. After preprocessing, the logs reduce to 43 MB. A total of 453,953 pages were accessed by 9,112 unique hosts. After session extraction using the time-out heuristic with $\tau = 30$ mins, a total of 37,753 sessions were extracted. Before converting the original cowpaths to concept-based cowpaths, a session consisted of visits to 12 webpages on an average over a time span of 16 minutes. For concept-based cowpaths, the average session consisted of 7 meta-pages. The following shows some sample concept-based cowpaths from each of the 10 clusters obtained after hyper-graph partitioning of the min-max path similarity matrix using a linearly decreasing $\vec{w}$.

CLUSTER 0 : Users skimming through the site, specially articles
path : authors 8 articles 8
path : home 47 authors 20 articles 22
path : home 18 articles 18
path : home 18 articles 16 coffeehouse 17
path : home 15 articles 15

CLUSTER 1 : Users spending some time studying the site and its structure
path : home 256 misc 256
path : home 37 misc 449 home 39 misc 175
path : home 271 misc 271
path : home 340 misc 194
path : home 101 misc 101

CLUSTER 2 : Users interested in authors and their articles
path : authors 138 home 138
path : home 109 authors 104 articles 106.5
path : home 180 authors 37 articles 108.5
path : authors 6 articles 6 authors 8 articles 76 authors 8 home 13
path : home 6 authors 159

CLUSTER 3 : Users following directions to articles and reading them seriously
path : misc 886 articles 1713 misc 1299.5
path : home 63 articles 615
path : home 4 misc 11 articles 1849
path : home 1182 misc 69 authors 50 misc 629 home 419 misc 1627 home 201 articles 282 misc 10 articles 802 home 173 misc 13 articles 792 misc 441 articles 427 home 12 misc 15 articles 346 misc 6 articles 1261 **path** : home 323 articles 24 authors 45 articles 1290

CLUSTER 4 : Users who follow directions to articles but skim through them
path : home 9 authors 5 articles 40 authors 35 articles 397 misc 15 authors 71
path : home 191 authors 76 articles 144 authors 29 articles 156 books 66
path : home 20 misc 37 articles 287 misc 10 articles 271 home 10 articles 20 home 23
path : home 6 authors 24 articles 234 authors 13 articles 35 authors 62.4
path : home 46 articles 267 authors 20 articles 83.25

CLUSTER 5 : Users surfing all over the place
path : home 20 coffeehouse 51 wo-men 239 personal 77 biztech 49 philosophy 55 movies 238 philosophy 191 books 12 creative 106 fun 26 wo-men 16 home 29
path : home 102 philosophy 56 wo-men 24 coffeehouse 46 fun 53 creative 30 personal 11 books 7 movies 34 contests 575 wo-men 35
path : fun 107 coffeehouse 11 wo-men 1 personal 3 philosophy 5 movies 4 books 3 contests 344 biztech 29
path : home 8 philosophy 31 personal 2 wo-men 2 coffeehouse 2 fun 5 creative 13 movies 1878 philosophy 1232 home 17 contests 43 coffeehouse 4 philosophy 1 movies 3 personal 2 contests 1 creative 3 fun 81 biztech 13 home 123
path : home 6 coffeehouse 753 wo-men 8 personal 3 philosophy 9 fun 113 creative 46 contests 2080 home 7 coffeehouse 4 wo-men 27 personal 32 philosophy 2 fun 3 creative 3 books 35 movies 8 contests 1035 home 86 coffeehouse 6 contests 327 home 13 contests 3630

coffeehouse 60 wo-men 4 personal 1 philosophy 2 fun 11 creative 3 books 3 movies 2 contests 43

CLUSTER 6 : Users who spend time in wo-men
path : home 63 wo-men 106.2
path : home 16 wo-men 215 personal 67 misc 18.625
path : wo-men 1159 fun 1159
path : home 4 wo-men 754 coffeehouse 774
path : home 20 fun 8 wo-men 1689 fun 9 home 356 creative 99

CLUSTER 7 : Users who spend time in personal
path : personal 1464 wo-men 176 home 151 personal 293
path : movies 914 personal 272.875
path : personal 384 home 49 personal 128 misc 113 articles 3 personal 2333.1
path : personal 1473
path : home 5 movies 435 personal 15 creative 81 personal 131 movies 18 coffeehouse 121

CLUSTER 8 : Users who spend time in home → coffeehouse
path : home 54 coffeehouse 1213.35
path : home 825 coffeehouse 32 wo-men 301 personal 71 coffeehouse 8 home 137
path : home 774 coffeehouse 278.7
path : home 26 wo-men 22 coffeehouse 665 fun 105 personal 13 coffeehouse 56 fun 43
path : personal 1428 coffeehouse 714

CLUSTER 9 : Users who have bookmarked coffehouse
path : coffeehouse 105 home 52.5
path : coffeehouse 282 home 15 articles 99
path : home 11 coffeehouse 93 authors 52 coffeehouse 340
path : coffeehouse 116 books 50 contests 55
path : coffeehouse 20 wo-men 33 coffeehouse 19

One notes that we have found really meaningful clusters. Varying the resolution of the combined similarity by varying $\vec{w}$ and varying the number of clusters generated leads to interesting results.

VIII. Concluding Remarks

The area of web-log mining, though growing rapidly, is still in its infancy [7]. The importance of proper preprocessing and filtering of server-logs was emphasized in [4] and was reinforced by our experience. Current approaches to cowpath clustering using correlations/associations as well as Markov-like approaches fail to use time information, and do not scale well with longer paths. A novel and noteworthy approach to this problem is taken by the Web Utilization Miner (WUM) system [8], in which individual paths are combined into an aggregated tree, and queries corresponding to desired path patterns are mapped onto the intermediate nodes of this tree structure. Again time information is not used.

In the paper, by defining a suitable similarity measure between any pair of sessions, we map the sessions into a similarity space, where efficient graph-partitioning algorithms can be used to do the clustering. This technique enables one to naturally incorporate both path and time information via the similarity metric. Moreover, it avoids curse-of-dimensionality problems encountered in traditional clustering methods (K-means etc.) applied to high dimensional spaces. The superiority of graph-partitioning methods for clustering has also been demonstrated in other web analytics domains such as clustering of web documents [9]. The second major contribution of this paper is the use of metadata such as pre-existing category labels. This captures the notion of *content* at a more abstract level. It is possible that extracting more detailed information from the contents of the pages being viewed can further enhance the quality of user clustering, provided the additional information complexity can be managed properly.

Acknowledgements

This research was supported in part by the NSF under Grant ECS-9000353, and by a gift from Neonyoyo Inc. We want to express our gratitude to Satya Prabhakar and Sangeeta Kshetry for providing the sulekha weblog data and the Department of ECE, UT Austin, for providing the departmental weblog data. All processing was done in a user-anonymous fashion to preserve the privacy of the visitors to the web-sites.

References

[1] Y. Fu, K. Sandhu, and M. Shih. A generalization-based approach to clustering of web usage sessions. In M. Spiliopoulou B. Masand, editor, *Web Usage Analysis and User Profiling*, pages 21–38. Springer, 2000.
[2] M. Chen, J. S. Park, and P. S. Yu. Data mining for path traversal patterns in a web environment. In *Proc. 16th Intl. Conf on Distributed Computing Systems*, pages 385–392, 1996.
[3] C. Shahabi, A. M. Zarski, and V. Shah J. Adibi. Knowledge discovery from users web-page navigation. In *Proc. 7th Intl Conf on Research Issues in Data Engg*, pages 20–29, 1997.
[4] R. Cooley, B. Mobashar, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web. In *Proc 9th IEEE Intl. Conf. Tools with AI (ICTA'97)*, Nov 1997.
[5] http://awsd.com/scripts/weblog/index.shtml.
[6] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392, 1998.
[7] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: Discover and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.
[8] M. Spiliopoulou and L. C. Faulstich. WUM: A tool for web utilization analysis. In *Extended version of Proc. EDBT Workshop WebDB'98*, pages 184–203. Springer Verlag, 1999.
[9] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. Impact of similarity measures on web-page clustering. In *Proceedings of Seventeenth National Conference on Artificial Intelligence : Workshop of Artificial Intelligence for Web Search (AAAI 2000), 30-31 July 2000, Austin, Texas, USA*, pages 58–64. AAAI, July 2000.