Self-Organizing Map based Web Pages Clustering using Web Logs

Dehu Qi
Computer Science Department
Lamar University
PO Box 10056
Beaumont, TX 77710, USA
dqi@cs.lamar.edu

Chung-Chih Li
School of Information Technology
Illinois State University
Campus Box 5150
Normal, IL 61790, USA
cli2@ilstu.edu

Abstract

A Web-based business always wants to have the ability to track users' browsing behavior history. This ability can be achieved by using Web log mining technologies. In this paper, we introduce a Self-Organizing Map (SOM) based approach to mining Web log data. The SOM network maps the web pages into a two-dimensional map based on the users' browsing history. Web pages with the similar browsing patterns are clustered together. Together with associate rules, the cluster generated by the SOM network has significant meaning to web browsing behavior. The experimental results demonstrate the feasibility and the effectiveness of this approach.

1 Introduction

There is an exponential increase of data available on the Web. The number of pages available on the Web is currently around 1 billion and is increasing at the rate of approximately 1.5 million per day. The Web-based business has been a key driving force for this rapid growth of the Web. Retailers on the Web need the ability to track users' browsing behavior history, which can increase the sale and build a strong customer relationship. This ability also can personalize the retailer's Web pages for different individual customers.

Although Web log mining is a relatively new field, it has generated a lot of interest and research in the past ten years. As a sub research field of Web Usage Mining, Web log mining is the process of applying data mining technologies to discover usage patterns from the Web data. One important source to discover such patters is the Web log data that contains users Web browsing history. Most of Web log data is generated automatically by Web servers.

Web mining mainly studies users' behavior of accessing and using the information on the Web [6] dynamically. Web usage mining deals with the analysis of Web usage patterns, for examples, user access statistical properties [12, 8, 2, 3], association rules and sequential patterns in user sessions [1, 6], user classification and Web page clusters based on user behavior [10, 7, 9, 14]. The results of Web usage mining can be used to understand user patterns in browsing information as well as to improve the accessibility of Web sites.

Kohonen et al. [5] used the SOM network to organize Web documents into a two-dimensional map based on the contents of documents. Web documents similar in content are located in similar regions on the map. This method automatically organizes the documents into meaningful clusters. Furthermore, the region, which represents node, indicates the similarity of the documents represented. Su et. al. [13] clustered Web documents using a recursive density based algorithm that can adaptively change its parameters intelligently. Smith and Ng [11] used a LOGSOM system to cluster Web pages using Web logs. The clustering is based on users' browsing history instead of the contents of Web pages. The system provided a visual tool to demonstrate the relationship between Web pages.

In this paper, we discuss a Self-Organizing Map (SOM) approach for the Web log mining. Different from other content-based Web page clustering approaches [5, 13, 14], the SOM-based approach clusters Web pages based on users' browsing patterns. The goal of this study is to test the feasibility of the SOM-based approach on Web page clustering using Web logs. The paper is organized as follows: Section 2 discusses related algorithms. Section 3 discusses the approach and methods. Section 4 presents the experimental results. Section 5 concludes our investigation.

2 Related Algorithms

2.1 SOM

Teuvo Kohonen [4] introduced the SOM network that reduced the dimensions of data through the use of self-organizing neural networks. The SOM network produces a map of usually one or two dimensions which plot the similarities of the data by grouping similar data items together. This mapping process reduces the problem dimen-

sions. The SOM network integrates dimensions reducing and clustering in one network. Figure 1 shows the mapping from a one-dimensional input to a two-dimensional array.

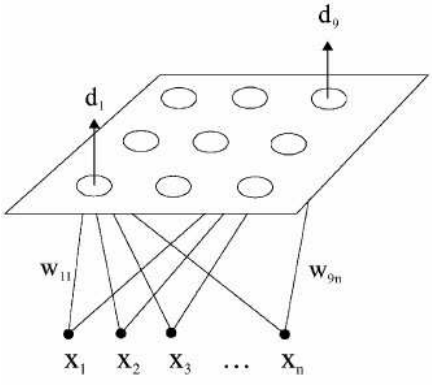

Figure 1: The Mapping from a one-dimensional input to a two-dimensional array [11].

The SOM network organizes itself by competing representation of the samples. Neurons are also allowed to change themselves in hoping to win the next competition. This selection and learning process makes the weights to organize themselves into a map representing similarities.

The algorithm of the SOM network is shown as follows:

1. Initialize Map
2. Set $t = 0$ and repeat the following steps until $t > 1$
   - Randomly select a sample
   - Get best matching unit
   - Scale neighbors
   - Increase t by a small amount
3. End for

The first step in constructing a SOM is to initialize the weight vectors. From there the algorithm selects a sample vector randomly and search the map of weight vectors to find the weight that can represent the sample best. Since each weight vector has a location, it also has neighboring weights that are close to it. The chosen weight is rewarded to perform better than a randomly selected sample vector. In addition to this reward, the neighbors of the weight are also rewarded. From this step we increase t some small amount because the number of neighbors and how much each weight can learn decreases over the time. This whole process is then repeated a large number of times, usually at least 1000 times.

The main advantage of using the SOM network is that SOM automatically (self-organizing) clusters documents. The SOM network also can be applied to a large scale of data.

2.2 K-Means

The *k*-means algorithm was introduced by J. MacQueen, and it had been one of the most popular clustering algorithms. This clustering algorithm represents each of *k* clusters C_j by the mean (weighted average) c_j of its point (called centroid). It initially selects clusters such that points are mutually farthest apart. Next, it examines each point and assigns it to one of the clusters depending on the minimum distance. The centroid's position is recalculated every time a point is added to the cluster. This process will be repeated until all the points are grouped into the *k* clusters. However, this algorithm does not work well if there are large differences in the data set. The equation for k-means algorithm is in Equation 1 and 2.

$$u_{ij} \in U_{c\times n}; Ci = \frac{1}{n_i}\sum_{j=1}^{n} X_j \tag{1}$$

$$\min J = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\left\|X_j - C_i\right\|^2 \tag{2}$$

In equations (1) and (2), X_j represents each point *j*'s co-ordinates and u_{ij} represents the hypothetical belonging of point *j* into cluster i (i.e., u_{ij} = 1 if *j* belongs to cluster *i*; u_{ij} = 0 if *j* belongs to any other cluster different from *i*).

3 SOM-based Web page clustering

3.1 Overall Architecture

Generally, our approach can be divided into three steps: data preprocessing, Web page mapping, and clustering analysis. Figure 2 shows these three steps.

In the data preprocessing step, a couple of methods are used to identify users, sessions, and transactions. The Web site topology is also identified in this step. In general, in this step, the raw Web data should be preprocessed into data abstractions for further processing.

After the data preprocessing step, SOM is used to cluster pages from similar navigating patterns. Unlike other Web personalization systems that usually find pages belonging to the same cluster based on the contents of the pages, our approach uses the user's current navigation pattern. Moreover, our SOM network uses the k-means clustering algorithm where more than one cluster will be considered at the same time for further analysis.

In the clustering analysis step, results from the Web page mapping step are stored in two-dimensional arrays. The Web site topology we identified in the preprocessing step will be used to filter patterns containing pages of a certain usage type. Clustering analysis can help the developer to get user's Web browsing patterns and predict the users' move when they brows some particular sites.

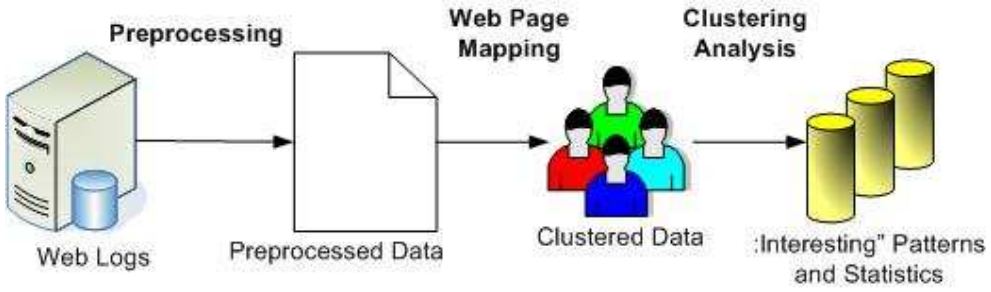

Figure 2: The Mapping from a one dimension input to the two-dimensional array [11].

3.2 Data Preprocessing

There are several pre-processing tasks to be done before executing the data mining algorithms on the Web server logs. These processes include data formatting, user identification, session identification, and transaction identification. The original server logs are formatted and grouped into meaningful transactions before being processed by the mining system. We describe each of these processes in the following paragraphs.

Data formatting The access log is saved to keep a record of every request made by the users. Since our main purpose is to facilitate more effective and efficient navigation, we only want to keep the log entries with information relevant to our purpose of organizing the Web pages. Some irrelevant log entries are deleted from the log file.

Sometimes a user requests a page that does not exist. This will create an error entry in the log. Since we are organizing the existing Web URLs, we are not interested in this kind of error entries, and hence these error entries shall be deleted. A users request to view a particular page often results in several log entries because the page consists of several materials such as graphics or small applets. However, we are only interested in, and hence only keep, what the user explicitly requests because we intend to design a system that is user-oriented.

User identification The task of identifying unique users is greatly complicated by the existence of local caches, corporate firewalls, and proxy servers. Therefore, some heuristics are commonly used to help identify unique users. We use the machines IP addresses to identify unique users.

User-session identification For logs that span a long period of time, it is very likely that different users may use the same machine to access the server Web sites. Therefore, we differentiate the entries into different user-sessions through a session timeout. That is, if two time stamps between page requests exceeds a certain limit, we assume the pages are requested by two different user-sessions, even though the IP address is the same.

Transaction identification The transactions are identified using maximal forward references. Each time a backward reference is made, a transaction is identified. A new forward reference indicates the next transaction for that session.

3.3 Web Page Mapping

K-Means Clustering After the user sessions and transactions are identified, we make a two-dimensional array in which each row is arranged for a transaction and each column is for a URL. Initially, the URLs that appear in a transaction are set to one in the corresponding row, and rest values are set to zero.

Initially, k transactions are selected at random for the k clusters. Then the means of the k clusters will be calculated. Afterwards,the distance between every transaction and the k clusters is calculated using the means of the k clusters. A transaction will be grouped into the cluster to which the distance is the shortest.

For each of these k clusters, we sum up the values of each column and calculate its new mean. The mean values are used as the weights for the groups, which are used to indicate the similarity between groups. The algorithm will be repeated until the weights become stable.

SOM The k groups of transactions and the set of unique URLs are the input to the SOM network. The input is represented by a two-dimensional m by k matrix, where m is the number of unique URLs and k is the number of transaction groups.

4 Experimental Results

We used Web log file for October, 2006 from the http://cs.lamar.edu as our test data. The data size is about 30MB with about 300,000 entries. Table 1, 2 and 3 shows the example of user identifications, session identifications, and transaction identifications.

The number of unique URLs generated by preprocessing is 188. We used a fixed value of 20 as the number of clusters, so the input to the SOM network is a 188 by 20 array. We have tested different parameters for the SOM network as follows: α varies from 0.2 to 0.9 and ω

Users	Browsing History
User 1	1-3-4-8-12-15
User 2	1-9-10
User 3	1-2-5-6-7-11-13-14

Table 1: User Identification

Users	Browsing History
User 1 Session 0	1-3-4-8
User 1 Session 1	12-15
User 2 Session 0	1-9-10
User 3 Session 0	1-2-5-6-7
User 3 Session 0	11-13-14

Table 2: Session Identification

varies from 1 to 40 where α represents the learning rate and ω determines the number of times a URL being presented within one learning cycle before the neighborhood size is decreased. In our algorithm, there are 18 learning cycles for organizing the Web pages. In particular, we decreased the neighborhood size from its initial value of 17 to 0. Table 4 and 5 shows the SOM map with (α = 0.1, ω = 40) and (α = 0.5, ω = 40), respectively.

From our experimental results, we find that, with $\omega = 40$, the two-dimensional array maps display clearest contesting. Table 6 shows part of the clusters with α = 0.1 and ω = 40. The SOM mapping self cluster the web page without prior knowledge.

To assess the effectiveness of our approach, we inspected the SOM map. We find that the approach indeed results a very meaningful SOM network in the sense that the Web pages are organized into clusters based on the similarity of their usage. Within a cluster, we can see that users are indeed likely to navigate Web pages within the same node, even though the SOM was given no information about the directory structure of the server and the contents of the Web pages. The SOM network has placed Web pages together when they are commonly accessed by the users in the same transactions.

Although it has been proven that clustering Web pages based on their contents is very effective and useful, it may be more advantageous to organize the Web pages in a user-pattern-based clustering. In such a way, the Web pages are organized for humans to search in a more effective and efficient manner due to its simplicity. Analysis the usage patterns of Web users can play an important role in assisting other users.

Users	Browsing History
User 1 Transaction 0	1-3-4
User 1 Transaction 1	1-3-8
User 1 Transaction 2	12-8-15
User 2 Transaction 0	1-9-10
User 3 Transaction 0	11-14
User 3 Transaction 1	1-2-5-6
User 3 Transaction 2	1-2-5-7
User 3 Transaction 3	11-13

Table 3: Transaction Identification

Cluster Number	Web Pages
8	901 902 903 904 905
10	8 21 23 89 90 133 134 136 168 180 284 285 286 288 289 290 313 319 328 337 338 343 344 351 357 359 374 392 393 394 399 406 410 416 421 434 442 448 454 455 456 466 480 487 498 499 513 583 593 789 1212 1230 1292
11	88 92 130 132 135 141 166 167 177 179 190 191 194 202 211 212 213 303 320 321 325 326 339 342 345 356 368 384 385 386 387 388 391 397 398 403 404 405 409 411 412 413 415 417 418 420 422 427 430 431 432 433 446 447 449 451 452 453 479 490 496 497 503 508 512 515 530 788 846 978 1213 1229 1259 1260 1288 1291 1293 1342
13	127 151 155 156 159 231 232 279 280 281 291 307 308 323 348 360 381 382 383 389 390 441 814 1273 1274 1275 1276 1277 1278 1286 1287 1289

Table 6: Part of clusters with α=0.5 and ω=40

5 Conclusions

We introduced a Self-Organizing Map (SOM) approach to the study of mining Web log data. Starting from the raw Web log data that is available in any Web server, we preprocessed it into distinct user transactions. We used the classical k-means algorithm to classify the URLs into clusters based on users' browsing history. The experimental results based on the data from the Web log of the server of our CS department demonstrate that our approach is very useful is a specified domain. The results of the clusters generated form the SOM network shows that our approach can effectively discover usage patterns. Our results can also be used to predict the user's browsing behavior based on the past experience.

0	2	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
0	3	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	2	0	5	2	0	0	0	0	1	0	3	0	0	0
0	8	0	1	0	2	0	0	0	33	0	0	0	2	4	1	0	0
0	2	0	0	1	3	0	0	22	0	0	0	0	0	0	0	0	0
0	0	0	0	2	0	6	0	0	0	0	0	0	2	0	0	0	0
0	1	2	1	1	1	3	0	0	1	0	0	0	3	0	0	0	0
0	0	0	0	1	4	6	0	12	1	1	119	0	0	0	0	1	0
2	0	0	0	0	0	0	2	0	1	0	1	0	0	0	1	0	1
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4: SOM map with α=0.1 and ω=40

0	2	2	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
0	0	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
0	0	116	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0
0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	35	3	2	0	1	3	0	0	1	0	0	0	0	1	3	0	0
0	2	0	0	0	1	0	2	0	0	0	0	0	0	3	1	1	0
24	12	0	0	0	0	0	0	0	0	0	0	0	0	4	0	1	0
1	4	6	0	16	3	0	0	0	0	0	0	0	0	0	0	2	0
0	2	6	0	0	5	2	1	0	0	0	0	0	0	0	0	0	0
2	1	2	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 5: SOM map with α=0.5 and ω=40

Acknowledgements

This work was supported by the Research Enhancement Grant of Lamar University.

References

[1] M. Baglioni, U. Ferrara, Romei A., S. Ruggieri, and F. Turini. Preprocessing and mining web log data for web personalization. In *the 8th Natational Conference of the Italian Association for Artificial Intelligence*, 2003.

[2] X. Huang, F. Peng, A. An, and D. Schuurmans. Dynamic web log session identification with statistical language models. *Journal of the American Society for Information Science and Technology*, 55(14):1290–1303, 2004.

[3] X. Jin, Y. Zhou, and B. Mobasher. Web usage mining based on probabilistic latent semantic analysis. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Seattle, WA, 2004.

[4] T. Kohonen. *Self-Organization and Associative Memory*. Springer-Verlag, New York, 1988.

[5] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela. Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585, 2000.

[6] J. Liu, S. Zhang, and J. Yang. Characterizing web usage regularities with information foraging agents. *IEEE Transactions on Knowledge and Data Engineering*, 2004(16):566–584, 2004.

[7] Rosa Meo, Pier Luca Lanzi, Maristella Matera, and Roberto O Esposito. Integrating web conceptual modeling and web usage mining. In *Proceedings of the sixth WEBKDD workshop: Webmining and Web Usage Analysis*, pages 105–115, Seattle, WA, 2004.

[8] B. Mobasher, H. Dai, and M. Tao. Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery*, 6:61–82, 2002.

[9] M. Nakagawa and B. Mobasher. A hybrid web personalization model based on site connectivity. In *Proceedings of the International Workshop on Web Knowledge Discovery and Data Mining*, pages 59–70, 2003.

[10] O. Nasraoui and C. Petenes. Combining web usage mining and fuzzy inference for website personalization. In *Proceedings of the International Workshop on Web Knowledge Discovery and Data Mining*, pages 37–46, 2003.

[11] Kate A. Smith and Alan Ng. Web page clustering using a self-organizing map of user navigation patterns. *Decision Support Systems*, 35(2):245–256, 2003.

[12] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.

[13] Zhong Su, Qiang Yang, Hong-Jiang Zhang, Xiaowei Xu, and Yu-Hen Hu. Correlation-based document clustering using web logs. In *34th Hawaii International Conference On System Sciences*, pages 5022–5027, Hawaii, 2001. IEEE Computer Society.

[14] A. Ypma and T. Heskes. Categorization of web pages and user clustering with mixtures of hidden markov models. In *Proceedings of the International Workshop on Web Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002.