A Novel Two-Phase SOM Clustering Approach to Discover Visitor Interests in a Website

Ahmad Ammari and Valentina Zharkova

University of Bradford – School of Computing, Informatics and Media
{A.N.Ammari, V.V.Zharkova} @Bradford.ac.uk – Bradford, United Kingdom

Abstract. Mining content, structure and usage data in websites can uncover browsing patterns that different groups of Web visitors follow to access the subjects that are truly valuable to them. Many works in the literature focused on proposing new similarity measures to cluster Web logs and detect segments of browsing behaviors. However, this does not reveal which contents the visitors are interested in since a Web page may contain many different topics. In this paper, a novel two-phase clustering approach based on Self Organizing Maps (SOM) is proposed to address this problem. A systematic process to prepare Web content data for clustering is also described.

1 Introduction

Mining the browsing behavior of website visitors is an important means to enhance its hyperlink structure to facilitate information access as well as to improve its contents to maintain visitor loyalty and attract additional Web surfers. The visitor interactions with the Web pages of a website are recorded by that website server in Web log files. Web usage mining [11] takes the data in these Web log files as input for data mining models. However, such log files usually contains millions of raw data records that requires a systematic process of data preparation to be suitable for these data mining applications [10].

In this work, a two-phase approach to segment the visitors of a University website based on their browsing behavior as well as on the content of that website is proposed. In particular, this approach firstly clusters the accessed Webpages based on their textual contents to form groups of similar pages, where each group corresponds to a certain topic or 'interest' that some visitors will be keen to find and browse on the website. Secondly, the visitor browsing sessions for the website are identified from the Web log files and clustered based on the time that they spend on each of the interests that have been determined by the first clustering phase. Because the data recorded in Web logs do not reveal the time spent on each interest by each visitor session, but only indicates the time spent on each Webpage in that website, a mapping process should be made to convert these visitor sessions from page-oriented to interest-oriented before applying the second clustering phase. Self Organizing Maps (SOM) is the clustering algorithm of choice for this work based on its popular success and effectiveness observed by related works such as [4] and [5] when working with high dimensional data such as textual data and Web log sessions.

A Summary of related work to this paper is provided in Section 2. In Section 3, the data preparation process for Web page contents is described in steps. An explanation of the proposed two-phase clustering approach is found in Section 4.

Experimental results are discussed in Section 5. Finally, conclusions for this work are drawn in Section 6.

2 Related Work

Mining Web data is usually categorized to three areas based on the type of data being mined: Web Content, Web Structure, and Web Usage Mining [2]. Web clustering is an unsupervised Web mining technique used to discover natural groupings among either Web pages or Web users that are homogeneous within each group but heterogeneous between different groups. Clustering Web data has been used in recent works such as in [1] and [7]. However, the first study used clustering on Web log files (Web usage data) to reveal the browser patterns of the website visitors. Unfortunately, this is insufficient to be able to discover the interests of visitors in the Web pages they access because log files only allow the identification of the pages that are visited, not the actual content and topics each page covers. On the other side, the study in [7] focused on clustering the Web pages themselves, which are also insufficient to discover the browsing patterns of the users who visit these pages.

Other studies such as [3] and [4] integrated Web content with Web usage mining to develop a new similarity measure to group visitors based on their interests in the Web pages. However, no procedural mapping process from Web pages to topics or 'interests' has been applied. This means that the implementation of such approaches may reveal groups of visitors having similar interests in certain Web pages, but not in certain topics or interests. The drawback here is that the content of one Webpage may cover many topics. For example, a portal Webpage in a University website may contain a section about the latest events that are happening in the campus, another section about recent short courses, and a third section about certain academic staff members. Therefore, even with having final visitor clusters determined by the clustering model, it is still difficult to uncover the real topics that each cluster of visitors is really interested in.

The contribution of this work is to address the above problem using a two-phase clustering approach. The first clustering phase is applied so Web pages that are similar in content can be grouped together. The result is a number of clusters where each cluster corresponds to a distinct interesting topic that could attract certain visitors. The role of the second clustering phase is to discover groups of visitors where each group contains those visitors who tend to access certain topics in the same manner, that is, who have the same interests in the Website under experimentation.

3 Preparation of webpage content data

The extraction and reformatting of useful content from the hybrid mixture of elements in Web pages is a tedious and error-prone process but unfortunately has not been addressed deeply in Web data preparation works [13]. To address such a problem, the following approach is applied to prepare the content of the University website that this study has addressed:

1. Select top $\boldsymbol{n}$ frequently visited Web page URLs. Any page URL having views $\boldsymbol{v}$ within the range $2000 \leq \boldsymbol{v} \leq 10000$ during the studied period were selected for content parsing.
2. Filter the selected URLs by discarding those referring to removed pages, multimedia objects (images, video files … etc), and unauthorized access.
3. Map each URL of a determined web page j to a unique number, j ∈ {1, 2… D}, where D is the total number of determined pages.
4. Parse the content of pages referred to by the filtered URLs into one page collection. A Perl HTML parsing tool was created and used.
5. Apply text filters to remove unneeded elements such as markup tags, spacing and line breaks between the markup elements and their attributes, embedded client-side scripts, forms, frames, and multimedia objects.
6. Remove stop words, which are irrelevant terms with respect to the main subject of the pages, such as determiners, conjunctions, and prepositions.
7. Stem words to their original roots. Porter's stemming algorithm [9] is used due to its simplicity and high performance.
8. Formulate the term – page matrix (S χ D) with each page j ∈ {1, 2… D} is represented as a vector of the weights of the relevant stem words i ∈ {1, 2… S} in the page collection, where each weight *ωij* for the i^{th} stem word in the j^{th} page is determined using the TF-IDF weighting scheme:

$$\omega_{ij} = \frac{freq_{ij}}{\max\limits_{1 \leq k \leq S} freq_{kj}} \times log \frac{D}{n_i} \qquad (1)$$

   $freq_{ij}$ is the number of occurrences of stem word *i* in web page *j* and n_i is the total number of occurrences of the stem word *i* in the web page collection.
9. Strengthening the term – page matrix by removing outlier stem words. Outlier stem words could be classified into three subtypes:
   a. Stem words that have too low weight values.
   b. Stem words that have too high weight values.
   c. Stem words that their number of non-zero weights in the page collection is too low.
10. Remove outlier Web pages, which have relatively little number of non-zero weights of stem words in their weight vectors, thus they tend to increase the sparseness of data and reduce the self-descriptiveness of pages in the collection.

4 Discovering Visitor Web Interests

4.1 Phase No. 1: Clustering Web Content

In order to present user-oriented Web content, it is very important to understand Web user segmentation based on their interests, not on the pages themselves, but on the topics these pages cover. To obtain such segmentation, Web content clustering is firstly applied on the term – page matrix for content clusters identification. Since each page is thought of as a vector in the dimensional space, it is wise to map the cosine of the angle between two page vectors, ***P**$_j$* and ***P**$_k$*, into a distance function ***Dis*** to be fed into the SOM algorithm.

$$Dis\left(P_j, P_k\right) = 1 - \frac{P_j \cdot P_k}{\left|P_j\right| \cdot \left|P_k\right|} = 1 - \frac{\sum_{i=1}^{S}(\omega_{ij} \cdot \omega_{ik})}{\sqrt{\sum_{i=1}^{S}(\omega_{ij})^2 \cdot \sum_{i=1}^{S}(\omega_{ik})^2}} \qquad (2)$$

Another popular distance measure usually used in SOM modeling is the Euclidean distance [6]:

$$Dis\left(P_j, P_k\right) = \sqrt{\sum_{i=1}^{S}(\omega_{ij} - \omega_{ik})^2} \qquad (3)$$

4.2 Mapping Visit Page Sessions to 'Visit Interest' Sessions

After assigning each page to the cluster that it mostly belongs to, a visit cluster session table can now be formulated, as shown in Table 1. This table can then be used to create the final input vectors, shown in Table 2, used to train the second SOM clustering phase, in which detected clusters are assigned interest weights based on the total time spent on each of them during each session. For example, in the sample shown in Table 2, the visitor in the first session spent a total of 56 and 240 seconds on visiting Web pages that belong to clusters B and C, respectively. The visitor in the second session spent a total of 96, 300, and 200 seconds on visiting pages that belong to clusters A, B, and D, respectively. These time totals are thus considered weights that represent the visitor interests in each cluster, that is, in each content topic.

Session ID	**Page ID**	**Cluster ID**	**Visit Time (sec)**
1	9	B	56
1	20	C	120
1	35	C	120
2	44	A	96
2	9	B	100
2	8	B	200
2	10	D	200

Table 1. visit cluster sessions

Total time spent on clusters (sec)				
Session ID	Cluster A	Cluster B	Cluster C	Cluster D
1	0	56	240	0
2	96	300	0	200

Table 2. visit session vectors based on time-valued interests

4.3 Phase No. 2: Clustering Web Visit Sessions

To segment visitors based on their interests, the second SOM clustering phase is performed. Hamming distance [6] between two visit session vectors $\boldsymbol{US_j}$ and $\boldsymbol{US_k}$ is the measure of choice here due to the reduced dimensionality of the vectors to be clustered:

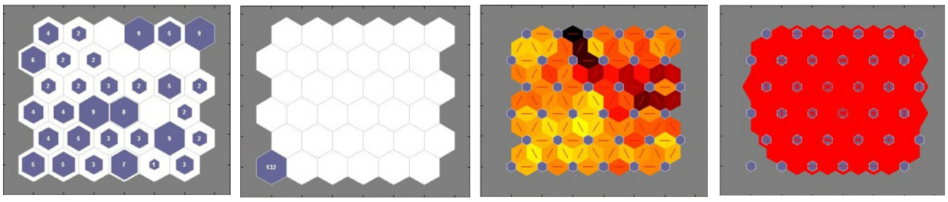
$$Dis\ (US_j, US_k) = \sum_{i=A}^{D} |us_{ji} - us_{ki}| \qquad (4)$$

$\boldsymbol{us_{ji}}$ and $\boldsymbol{us_{ki}}$ are the time weights of cluster $\boldsymbol{i} \in$ {A, B, C, D} in sessions $\boldsymbol{US_j}$ and $\boldsymbol{US_k}$, respectively.

5 Experimental Results

We began testing our framework by clustering the interesting Web pages in Bradford University School of Informatics website [12] because this site is sufficiently large in both the number of contained pages and the number of records in Web log files. The Matlab Neural Network Toolbox [8] was the platform of choice to train the SOM experimental models.

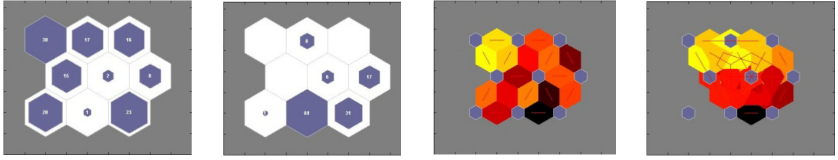
Fig. 1 shows the neuron classifications (a & b) and distances between neighbor neurons (c & d) for a 6 X 6 SOM layer trained using the *Euclidean* distance with a term-page matrix of 132 pages, each having 374 normalized stem word weights. While this SOM was able to recognize 3 to 4 page clusters in such a high dimensional space, another *cosine* distance SOM with same size clearly failed to recognize more than one cluster:



(a) Euclidean (b) Cosine (c) Euclidean (d) Cosine

Fig. 1. 6 X 6 SOMs trained using the *Euclidean* and *Cosine* distances

Fig. 2 shows another two smaller 3 X 3 SOM models trained with a 132 X 216 term – weight matrix after further filtering of noisy stem words. This has clearly resulted to an agreement on the number of Web topic clusters by the two distance measures:



(a) Euclidean (b) Cosine (c) Euclidean (d) Cosine

Fig. 2. 3 X 3 SOMs trained using the *Euclidean* and *Cosine* distances

6 Conclusion

A preparation process to the unstructured elements in web resources is indeed tedious, error-prone and should follow a systematic approach to derive the required input data set that will really make a difference in Web mining model building such as Web content clustering. Experimental results show that taking more care with textual filtration of high dimensional Web content as well as estimating the correct number of neurons in a SOM layer can result in more valid SOM clusters that more than one distance measure can agree on. This Web content SOM clustering is the first phase in an approach that aims to achieve a better understanding of the real interests that attract the visitors of a website and make them spend times of browsing its pages.

Future work will continue to investigate the efficiency of this novel approach through implementing the second clustering phase, discovering the distinct groups of the website visitors, and interpreting the browsing patterns in each group with further analysis. Other more recent text parsers, such as Stanford's parser, will be tried in the preparation step to detect and evaluate any possible text processing improvement.

References

[1] A. Joshi and R. Krishnapuram, On Mining Web Access Logs. In Proceedings of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 63-69, 2000.

[2] B. Berent, A. Hotho and G. Stumme, Towards Semantic Web Mining, *Proceedings of the First International Semantic Web Conference,* 264-278, Sardinia, Italy, June 9-12, 2002.

[3] B. Mobasher, T. Luo, Y. Sung, and J. Zhu, Integrating Web Usage and Content Mining for More Effective Personalization, *In Proceedings of the International Conference on E-Commerce and Web Technologies,* September, Greenwich, UK, 2000.

[4] J. D. Vel´asquez, H. Yasuda, T. Aoki, R. Weber: A new similarity measure to understand visitor behavior in a web site. In: *IEICE Transactions on Information and Systems*, E87-D(2), 389-396, 2004.

[5] J. D. Vel´asquez, H. Yasuda, T. Aoki: Using Self Organizing Feature Maps to acquire knowledge about visitor behavior. In: *Proceedings of the knowledge-based intelligent information and engineering systems*, pp. 951—958, Oxford, UK 2003.

[6] K. J. Cios, W. Pedrycz, R. W. Swiniarski, L. A. Kurgan: Data Mining, A Knowledge Discovery Approach. Springer Science Business Media, New York 2007.

[7] M. Perkowitz, AdaptiveWeb Site: Cluster Mining and Conceptual Clustering for Index Page Synthesis, *Dissertation for degree of Doctor of Philosophy, University of Washington,* 2001.

[8] Matlab Neural Network Toolbox, http://www.mathworks.co.uk/products/neuralnet/

[9] Porter, M.: An Algorithm for suffix stripping. Program, 14(3), 130-137, 1980.

[10] R. Cooley, B. Mobasher, J. Srivastava. Data preparation for mining World Wide Web browsing patterns. *Journal of Knowledge and Information Systems*, Vol. 1, pages 5-32*,* 1999.

[11] S. K. Pal, V. Talwar and P. Mitra, Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, *IEEE Transactions on Neural Networks*, Vol. 13, No. 5 pages 1163-1177, September, 2002.

[12] University of Bradford School of Informatics, http://www.inf.brad.ac.uk/home/

[13] Y. Li, C. Zhang, S. Zhang: Cooperative strategy for Web data mining and cleaning. In: *Applied Artificial Intelligence*, Vol. 17, pp. 443—460 2003.