A Comparison Framework of Similarity Metrics Used for Web Access Log Analysis

Yusuf Yaslan and Zehra Cataltepe

Istanbul Technical University, Computer Engineering Department, Maslak
34469 Istanbul , Turkey
{yyaslan,cataltepe}@itu.edu.tr

Abstract. In this paper, different types of web session similarity metrics are compared and combined for better web session clustering. Syntactic and co-occurrence information are used for similarity calculation. Syntactic information on a web page includes the place of the page in the directory hierarchy. Co-occurrence information is the amount of the occurrences of two web pages in the same sessions. Vector space representation of sessions and cosine, pearson and jaccard similarities between them are also used as a similarity metric.
Clustering quality is used as the goodness measure of a similarity. The clustering quality is given by the internal and external cluster similarity. First, clustering quality is evaluated when different similarity metrics (jaccard, pearson, cosine, syntactic, co-occurrence) are used. Similarity calculation is performed for different number of clusters from 1 to the number of sessions. For reasonable cluster numbers (15-100) syntactic similarity results in the best internal cluster similarity, followed by the co-occurrence similarity. Finally the best two methods and others are used for similarity combination. It is found that linear combination of similarity metrics decreases the external cluster similarity.
Similarity metrics are also evaluated using Hubert's statistics. It is found that syntactic similarity alone gives the best results according to Hubert's statistics followed by the linear combination of syntactic and co-occurrence similarity.

1 Introduction

Every day both the number of web pages on the web and the data produced by the people who browse those pages keep increasing. The vastness of the amount and usage opportunities of these data resulted in the web usage mining techniques[1]. Web usage mining is the processing of web usage data and it is used for customization for web users by using client-server transactions on one or more web servers. Hence it can be used for improving the web cache performance, personalize the browsing experience and improve search engines[3]. Previously in [4, 5] web session logs are used to better understand user behavior in order to design better environments and applications. The transactions are grouped by sessions where a session consists of transactions of a web user within a specific

time. In order to do customization for users, similarities between web sessions are computed. Then each session is assigned to a group by using the similarity value together with different clustering techniques such as graph partitioning.

In this paper syntactic, co-occurrence and vector space based features are obtained. Syntactic similarity gives information on the web site directory hierarchy. Co-occurrence information is the number of web pages visited together in sessions. Visited web sites information is used for constructing the session vectors. Cosine, Extended Jaccard and Pearson Correlation similarity metrics are used for obtaining vector space based similarities. Finally linear combinations of similarities are evaluated.

In [4, 5] multi-modal feature vectors are generated using the content, usage and topology of the analyzed web site. The content modality is the unique words that appear in the web page. Usage modality is obtained by tokenizing the URLs of the site by using the "/" delimiter. The other modalities are obtained by using inlinks and outlinks of the web page. Outlinks of a page represent the pages that are reachable from the page. Similarly inlinks of a page represent the pages that points to the web page. The topology of a site is represented using an adjacency matrix, where the rows correspond to the outlinks and columns correspond to the inlinks.

Previously in [2] cosine, pearson correlation and extended jaccard similarity metrics are used in conjunction with different clustering methods (random, hyper-graph partitioning, self organizing feature maps) for web page clustering. It is shown that cosine and extended jaccard similarities give the best results. In this paper we cluster not the web pages but the web sessions. In [2], similarities between TF.IDF vectors that correspond to occurrences of words in a document are used. In this study, we use similarities between sessions, i.e. each component of the vectors we use corresponds to a web page, as opposed to a word in a web page.

Syntactic similarity is calculated on the tokens of the visited URLs within a session. Previously in [6] syntactic similarities of web pages are combined using cosine similarity in order to obtain dissimilarity values for web log clustering. In this paper, we also combine different similarity metrics together with syntactic similarity for obtaining similarity values between web sessions.

The rest of the paper is organized as follows: In Section 2, the way we represent sessions and syntactic and co-occurrence similarities between them, together with different similarity metrics are introduced. In Section 3, details of combination methods of different similarity metrics are given. In Sections 4 and 5, we present the experimental results and the conclusions.

2 Session Similarity Metrics

2.1 Session Representation Using Vector Space Model

Each session consists of a sequence of visited web pages. In our vector space representation, we represent a session with a vector whose size equals the total

number of pages. For example, let the web site consist of a total of 5 web pages, represented using letters A through E. If a session consists of A → B → D, then vector space representation for the session could be: (1,1,0,1,0).

Different types of path weighting schemes can be applied[5]:

- Uniform: Each page has an equal weight in the session. e.g. A → B → D=(1,1,0,1,0)
- TF.IDF(Term Frequency Inverse Document Frequency): Each page receives a TF.IDF weighting. The number of a specific web page within a session is divided by the total number of occurrances of the web page in the logs.
- Linear Order (or Position): The web pages are weighted using the order of page accesses in the session e.g. A → B → D=(1,2,0,3,0).
- View Time: The web pages are weighted by using the time spent on that page accesses in the session e.g. A(10sec) → B(20s) → D(15s) = (10,20,0,15,0).
- Various Combined Weighting: Each page is weighted with various combinations of the TF.IDF, Linear Order, and View Time path weighting. Linear Order+View Time, e.g.: A(10sec) → B(20s)→ D(15s) = (10,40,0,45,0).

In this study, we have used linear order and view time schemes for web session similarity calculation.

We represent a session i using a vector $s(i)$:

$$s(i) = (s(i,1), s(i,2), s(i,3), ..., s(i,N_i)) \tag{1}$$

$s(i)$ is an N_i dimensional vector of web pages $s(i,j)$ where N_i is the number of pages visited during the session.

The first method uses a vector space model similar to the one in document categorization. If page k has been visited during session i, then the clickstream vector $\Theta(i)$ that contains the page visit information can be computed as:

$$\Theta(i) = \begin{cases} \Theta(i,k) = \text{number of k' } if \ \exists \ k' \ s.t. \ s(i,k') = p_k \\ \Theta(i,k) = 0 \qquad\qquad otherwise \end{cases} \tag{2}$$

$\Theta(i)$ is an N dimensional binary vector where N is the total number of pages on the web site we are interested in. p_k is the kth web page. The number of visit of a web page is represented in the vector model. Cosine, pearson correlation and extended jaccard similarity, which we all define below, work on the vector space model.

2.2 Cosine Similarity

Cosine similarity captures scale invariant understanding of similarity and it is defined related the angle between two vectors as follows[2]:

$$Sim_{cos}(s(i), s(j)) = \frac{\Theta(i)^T \Theta(j)}{\sqrt{\Theta(i)^T \Theta(i) \Theta(j)^T \Theta(j)}} \tag{3}$$

2.3 Pearson Correlation Similarity

Pearson correlation reflects the degree of linear correlation between two sessions:

$$Sim_{pear}(s(i), s(j)) = \frac{1}{2}\left(\frac{(\Theta(i) - \overline{\Theta(i)})^T(\Theta(j) - \overline{\Theta(j)})}{\sqrt{(\Theta(i) - \overline{\Theta(i)})^T(\Theta(j) - \overline{\Theta(j)})}} + 1\right) \tag{4}$$

where $\overline{\Theta}$ refers to the mean value of the feature vectors.

2.4 Extended Jaccard Similarity

Extended Jaccard similarity measures the length dependent measure of similarity [2]:

$$Sim_{jaccard}(s(i), s(j)) = \frac{\Theta(i)^T\Theta(j)}{(\Theta(i)^T\Theta(i) + \Theta(j)^T\Theta(j) - \Theta(i)^T\Theta(j))} \tag{5}$$

2.5 Syntactic Similarity

Syntactic similarity uses the dependencies between the web page directories. According to [6] syntactic similarity between k^{th} and l^{th} web pages of sessions $s(i)$ and $s(j)$ is calculated as follows:

$$Sim_u(s(i,k), s(j,l)) = min\left(1, \frac{|p(i,k) \bigcap p(j,l)|}{max(1, max(|p(i,k)|, |p(j,l)|) - 1)}\right) \tag{6}$$

where $p(i,k)$ denotes the path traversed from the root node to the node corresponding to the kth URL, and $|p(i,k)|$ indicates the length of this path. This similarity measures the amount of overlap between the paths of the two visited web pages.

The similarity between sessions $s(i)$ and $s(j)$ is calculated as:

$$Sim_{synt}(s(i), s(j)) = \frac{\sum_{k=1}^{N}\sum_{l=1}^{N}\Theta(i,k)\Theta(j,l)Sim_u(s(i,k), s(j,l))}{\sum_{k=1}^{N}\Theta(i,k)\sum_{l=1}^{N}\Theta(j,l)} \tag{7}$$

Note that if two sessions are identical then the similarity might be small depending on the number of pages accessed.

2.6 Co-Occurrence Similarity

Co-occurrence similarity is derived from document clustering[8]. The amount of occurrences of the two sessions are proportional to the minimum number of visits of two pages as follows:

$$Sim_o(s(i,k), s(j,l)) = \frac{\#(p(i,k) \bigcap p(j,l))}{min(\#(p(i,k)), \#(p(j,l)))} \tag{8}$$

where $\#(p(i,k) \bigcap p(j,l))$ refers to the amount of the k^{th} and l^{th} web pages visited together. Similarly $\#(p(i,k))$ refers to the total occurrence number of k^{th} web page.

Similar to the syntactic similarity, co-occurrence similarity between sessions $s(i)$ and $s(j)$ is calculated between all pages in two sessions as follows:

$$Sim_{cooc}(s(i), s(j)) = \frac{\sum_{k=1}^{N} \sum_{l=1}^{N} \Theta(i,k)\Theta(j,l) Sim_o(s(i,k), s(j,l))}{\sum_{k=1}^{N} \Theta(i,k) \sum_{l=1}^{N} \Theta(j,l)} \quad (9)$$

3 Combination of Similarity Metrics

In our study, we have combined cosine, pearson correlation and extended jaccard similarity values with syntactic similarity in order to obtain similarity between two sessions. This combination is based on the intuition that we may be able to gain information from vector and URL of the web pages which may carry information about the visited webpages.

3.1 Linear Combination

We have used linear combination. We have combined the best two results and all similarity values alone. Our aim is to gain balanced information from vector space model and URL based syntactic information. The linear combination of syntactic and co-occurrence similarity is obtained as follows:

$$Sim_{com} = \alpha Sim_{synt} + \beta Sim_{cooc} \quad (10)$$

where $\alpha + \beta = 1$. In our experiments we have used α and β as $(0.5, 0.5)$ pairs. The second combination is obtained using all similarity values as follows:

$$Sim_{com} = \alpha Sim_{synt} + \beta Sim_{cooc} + \gamma Sim_{cos} + \eta Sim_{pear} + \delta Sim_{jaccard} \quad (11)$$

where $\alpha + \beta + \gamma + \eta + \delta = 1$. In our experiments we have used α, β, γ, η and δ as $(0.2, 0.2, 0.2, 0.2, 0.2)$.

4 Experimental Results

In our experiments we used a real dataset obtained from web logs of a Turkish web portal. The portal provides e-mail and news as well as web search service. Different types of domains are available such as; news, sport, cinema, finance, TV, fortune, estate and etc. More than 28000 web sessions are obtained from web logs. In order use the web server logs, first data cleaning is applied. Polliwog 0.6 (http://polliwog.sourceforge.net) is used for web session log data cleaning. In our experiments, we only use sessions whose lengths are greater than or equal to 7 pages. After log cleaning we have obtained 1321 web sessions for clustering. In order to compare different similarity metrics and their combinations, we

have used Cluto (http://www-users.cs.umn.edu/ karypis/cluto) clustering program. The output of the program is internal cluster similarity, external cluster similarity and their standard deviations.

In order to compare the experimental results fairly we have obtained the mean internal cluster similarity versus the mean external cluster similarity plots. The means are computed using the internal and external similarity of each cluster weighed by the number of data points in the cluster. The experiments are performed for ten different cluster sizes: 1,5,10,15,20,50,100,500,1000 and 1321.

In figure 1 mean internal cluster similarity versus mean external cluster similarity values for syntactic, jaccard, pearson, cosine and co-occurrence similarity are given. Clusters numbers are in ascending order of the internal similarity.

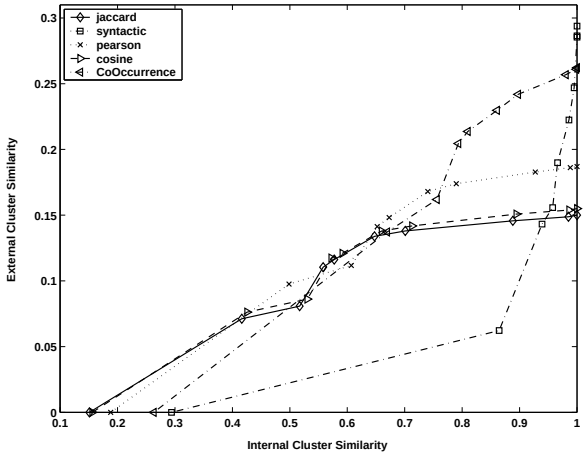

Fig. 1. Mean internal Cluster similarity versus mean external cluster similarity for 1-1321 clusters obtained by using different similarity metrics

4.1 Combining Similarity Metrics

In figure 2 mean internal cluster similarity versus mean external cluster similarity values for i) syntactic, ii) linear combination of syntactic, jaccard, pearson, cosine and co-occurrence similarity (linear comb.all sim. in figure) and iii) linear combination of syntactic and co-occurrence similarity (linear comb. of syntactic co-occurrence in figure) are given. Clusters numbers are again in ascending order of internal similarity. As it can be seen in figure 2 linear combination of all similarity metrics decrease the external cluster similarity when number of clusters are greater than 15.

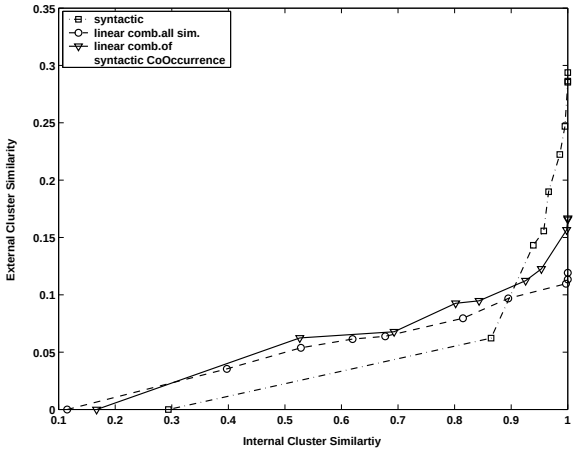

Fig. 2. Mean internal Cluster similarity versus mean external cluster similarity for 1-1321 clusters obtained by using linear combination of similarity metrics

4.2 Clustering Evaluation using Hubert's Γ Statistics

The goodness of a clustering can be obtained using the correlation between the similarity matrix and an ideal version of the similarity matrix based on clustering labels[9]. Ideal clustering can be defined as the clustering where internal similarities between objects in the clusters are 1 and external similarities are 0.An ideal similarity matrix for clusters can be obtained by sorting rows and columns of the similarity matrix. This ideal similarity matrix has a block diagonal structure. Thus intra-cluster similarity is 1 and all other entries are 0. Hubert's Γ statistics between similarity matrix S_{sim} and ideal similarity matrix S_{ideal} can be formulated as follows[10];

$$\Gamma = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} S_{sim}(i,j) S_{ideal}(i,j) \tag{12}$$

Where N is the dimension of the similarity matrix and M is the pairwise comparison and can be obtained as:

$$M = N(N-1)/2 \tag{13}$$

In figure 3 cluster number versus Hubert's Γ statistics for different similarity metrics are shown. As it is shown, the best Γ value is obtained for syntactic similarity followed by linear combination of syntactic and co-occurrence similarity. Note that if a web portal is hieratically well designed syntactic similarity enforces objects to have higher similarity values. Hence the Γ value for syntactic similarity overcomes the others. However the combination of similarity metrics not only decrease the external cluster similarity but also increases the Γ value.

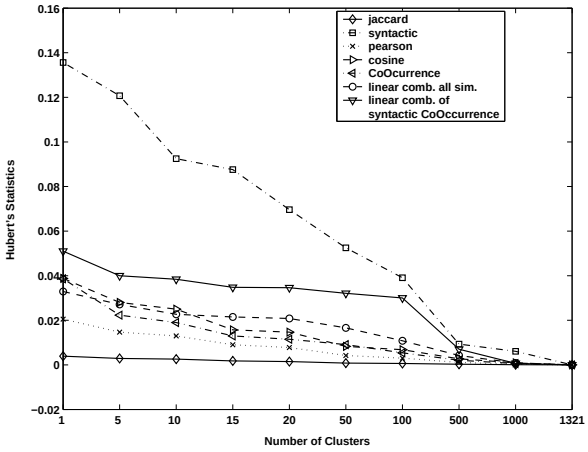

Fig. 3. Cluster number versus Hubert's statistics Γ for different similarity metrics

5 Conclusions

In this paper, different similarity metrics and their linear combinations are evaluated according to internal and external similarities and Hubert's Γ statistics. We have evaluated different α and β values for linear combination. However most of them gave similar results. Hence we give equal weight to all similarities (i.e. $\alpha = \beta = 0.5$). Experimental results have shown that, syntactic similarity gives the best Γ statistics because it enforces similarities to approach a maximum. However, combination of the syntactic and co-occurrence similarity metrics decreases the external similarity between clusters and increases the Γ statistics and gives satisfactory internal similarity results. Thus can be used for web session clustering for different numbers of clusters. Note that for syntactic similarity, increase in the cluster number increases the external cluster similarity.

Acknowledgements

This paper is partially supported by Turkish Scientific and Technical Research Foundation (TUBITAK) EEEAG Project number 105E162. Authors would like to thank Murat Goksedef and Nildem Demir of ITU for providing the user session data and Sule Gunduz-Oguducu and Sima Etaner-Uyar for useful discussions.

References

1. Pierrakos, D., Paliouras,G., Papatheodorou,C., Spyropoulos,C.D.:Web Usage Mining as a Tool for Personalization: A Survey,User Modeling and User-Adapted Interaction , 13,(2003) 311-372.

2. Strehl, A., Ghosh, J., Mooney, R. :Impact of Similarity Measures on Web-page Clustering,Workshop of Artificial Intelligence for Web Search, (2000).
3. Deshpande, M., Karypis, G. : Selective Markov Models for Predicting Web Page Accesses, ACM Transactions on Internet Technology, 4(2), (2004)163-184.
4. Heer,J. ,Chi, E. :Identification of Web User Traffic Composition using Multi-Modal Clustering and Information Scent, in Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining, Chicago, (2001),51-58.
5. Heer, J. ,Chi, E. :Mining the Structure of User Activity using Cluster Stability,in Proc. of the Workshop on Web Analytics, SIAM Conference on Data Mining, Arlington VA,(2002).
6. Joshi, A., Krishnapuram,R.:On Mining Web Access Logs,in Proc. of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, (2000),63-69.
7. Nasraoui, O., Krishnapuram, R. :An Evolutionary Approach to Mining Robust Multi-Resolution Web Profiles and Context Sensitive URL Associations,International Journal of Computational Intelligence and Applications, 2 (3),(2002),1-10.
8. Pal, S., Talwar, V.,Mitra,P. :Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions,IEEE Transactions on Neural Networks, 13,(5), (2002), 1163-1177.
9. Tan,P.N., Steinbach, M., Kumar, V. :Introduction to Data Mining,Addison-Wesley, (2005).
10. Theodoridis ,S., Koutroumbas ,K.:Pattern Recognition,Academic Press, (1999).