Merging Data Mining Techniques for Web Page Access Prediction: Integrating Markov Model with Clustering

Trilok Nath Pandey[1], Ranjita Kumari Dash[2] , Alaka Nanda Tripathy[3] ,Barnali Sahu[4]

[1] Siksha O Anusandhan University Department of Computer Science & Technology, Bhubaneswar, Orissa, India

[2] Siksha O Anusandhan University Department of Computer Science & Technology, Bhubaneswar, Orissa, India

[3] Siksha O Anusandhan University Department of Computer Science & Technology, Bhubaneswar, Orissa, India

[4] Siksha O Anusandhan University Department of Computer Science & Technology, Bhubaneswar, Orissa, India

Abstract

Web page access prediction gained its importance from the ever increasing number of e-commerce Web information systems and e-businesses. Web page prediction, that involves personalizing the Web users' browsing experiences, assists Web masters in the improvement of the Website structure and helps Web users in navigating the site and accessing the information they need. The most widely used approach for this purpose is the pattern discovery process of Web usage mining that entails many techniques like Markov model, association rules and clustering. Implementing pattern discovery techniques as such helps predict the next page to be accessed by the Web user based on the user's previous browsing patterns. However, each of the aforementioned techniques has its own limitations, especially when it comes to accuracy and space complexity. This paper achieves better accuracy as well as less state space complexity and rules generated by performing the following combinations. We integrate low -order Markov model and clustering. The data sets are clustered and Markov model analysis is performed on each cluster instead of the whole data sets. The outcome of the integration is better accuracy than the combination with less state space complexity than higher order Markov model.

Keywords: *Markov Model, Pattern discovery, clustering, space complexity, silhouette value.*

1. Introduction

It is worth canvassing the results of integrating Markov model with another prediction algorithm, clustering. Clusters are employed to guide the prediction system. They help predict the Web pages that are close to a user-requested page in a cluster model. Similar to the other prediction models, the cluster model tries to discover the statistical correlation between Web pages using Web access patterns mined from a Web log. However, prediction is performed on the cluster sets rather than the actual sessions. The main issue that affects the clustering accuracy is producing the selected features for partitioning. For instance, partitioning based on semantic relations hips or contents [7] or link structure [16] usually provides higher accuracy than partitioning based on bit vector, spent time, or frequency. However, even the semantic, contents and link structure accuracy is limited due to the unidirectional Nature of the clusters and the multidirectional structure of Web pages . We begin with the problem of focusing on integration process between markov model and clustering in section 2 and 3 we present and relate several important notions for session categorization in section 4 and 5. Finally we compare markov model, clustering and merging both and simulate our result. Then we end up with some conclusion in succeeding section.

2. Integration Process

The focus of this paper is on improving the Web page access prediction accuracy and state space complexity by combining Markov model [1] and clustering techniques [18],[19]. This paper explains the Markov model and clustering integration process.

3. Integration Algorithm

The training process takes place as follows:

1 Us e feature s election, allocate similar Web

sessions to appropriate categories.

2 Decide on a suitable k -means algorithm distance measure.

3 Decide on the number of clusters k and partition the Web sessions into clusters.

4 FOR each cluster

5 Return the data to its uncategorized and expanded state.

6 Perform Markov model analysis on each of the clusters.

7 ENDFOR

The prediction process [2] or test phase involves the following:

1 FOR each coming session

2 Find its closes t cluster

3 Use the corresponding Markov model to make prediction

4 ENDFOR

4. Feature Selection

The first step of the training process is feature selection and categorization. Since the improved Web personalization is subject to proper preprocessing of the usage data [7], [8]. It is very important to group data according to some features before applying clustering techniques. This will reduce the state space complexity and will make the clustering task simpler. However, failing to appropriately select the features would result in wrong clusters regardless of the type of clustering algorithm that is used. [15] presented methods aim at finding common categories among a set of transactions and mapping the trans actions to the predefined categories .

5. Session Categorization

Consider a data set D containing N number of sessions. Let W be a user session including a sequence of pages visited by the user in a visit. D= {W1… WN}.Let P = {p1, p2… pm} be a set of pages in a Web site. Since Markov model techniques will be implemented on the data, the pages have to remain in the order by which they were visited. Wi = ($p_1^i, \ldots, p_L^i$) is a session of length L composed of multivariate feature vectors p. The set of pages P is divided into a number of categories Ci where Ci = {p1; p2 …pn}. This results in less number of pages since Ci $\subset$ P and n < m. For each session, a binary representation is used assuming each page is either visited or not visited. If the page is visited, a weight factor w is added to the pages representing the number of times the page was visited in the new session Si. Si $\{(c_1^i, w_1^i), \ldots, (c_j^i, w_j^i)\}$. Ds is the data set containing N number of sessions SN. The categories are formed as follows:

Input: D containing N number of sessions WN.

(1)FOR each page pi in session Wi

(2) IF pi $\subset$ Ci

(3) wi.count++

(4) ELSE,

(5) wi = 0

(6) ENDIF

(7) ENDFOR

Output: Ds containing N number of Sessions SN.

5.1 Experimental Evaluation

5.1.1 Data Collection and Preprocessing

After Web session identification, session categorization took place and the details of the number of categories for each data set are represented in Table 5.1. After identifying all categories for each data set, it was necessary to run the session categorization algorithm.

Table 5.1: Number of categories

	D1	D2	D3	D4
# Sessions	2,520	4,356	13,745	5,673
# Categories	196	154	267	231

Table 5.2 below reveals part of session categorization implemented on data set (D2). The first row represents the category number and each row thereafter represents a session. For instance, the first session has 7 pages where three pages belong to category 5, one page belongs to category 7, two pages belong to category 10 and one page belongs to category 11.

Table 5.2: Session Categorization

1	2	3	4	6	8	9	10	15	19	23	26	30	34	50
0	0	0	0	3	0	1	0	0	2	1	0	0	0	0
0	0	0	1	0	0	0	0	5	0	1	0	0	1	0
1	0	0	0	0	4	0	0	0	0	0	0	2	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
1	0	0	2	0	1	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	1	2	0

This session categorization resulted in Web sessions of equal lengths [3] [4]. The extract in table 5.10 represents only around 10% of the actual categories. All categorized sessions were represented by vectors with the number of occurrence of pages as weights. This will draw sessions with similar pages closer together when performing clustering techniques. The next step before implementing

k-means clustering algorithm was to identify the number of clusters used and evaluate the most appropriate distance measure for all 4 data sets.

5.1.2. Number of Clusters (k)

Identifying the most appropriate number of clusters for all four data sets is a complex task because of lack of a one evaluation metric for the number of clusters. Different data sets with different number of categorized sessions' leads to different results according to different number of clusters. Generally speaking, larger data sets with more sessions are best clustered using more clusters than smaller data sets [10]. Therefore, the number of clusters used for each data set was a result of applying k-means algorithm to each data set and, then applying ISODATA algorithm to the resulting clusters. For instance, we achieved best results for D1 when k=7, for D2 when k=9,for D3 when k=14 and for D4 when k=10. This proves that a larger number of Web sessions is best clustered using a larger k. All clusters were attained using Cosine distance measure. Figure 5.1 depicts the 7 clusters of data set D1, Figure 5.2 depicts the 9 clusters of data set D2, Figure 5.3 depicts the 14 clusters of data set D3 and Figure 5.4 depicts the 10 clusters of data set D4.

5.1.3 Distance Measures Evaluation

Our basic motivation behind using clustering techniques is to group functionally related sessions together based on Web services requested in order to improve the Markov model accuracy. The Markov model accuracy increases if the Web sessions are well clustered due to the fact that more functionally related sessions are grouped together. To help find an appropriate *k*-means clustering distance measure we can apply to all four data sets, we examine the work presented by [11], [12]. In order to back up their findings, we calculate the entropy measures, we perform means analysis and we plot different clusters using different distance measures for data set D1. Table 5.3 lists entropy measures for only some of the clusters for data set D1 due to space limitation. The table demonstrates that, in general, Cosine and Pearson Correlation yield lower entropy measures and, therefore, they constitute better clusters than the other distance measures. Figure 5.5, Figure 5.6, Figure 5.7, Figure 5.8 and Figure 5.9 represent clusters using Euclidean, Hamming, City Block, Pearson Correlation and Cosine distance measures respectively for data set D1. They plot the silhouette value represented by the cluster indices displaying a measure of how close each point in one cluster is to points in the neighboring clusters. The silhouette measure ranges from +1, indicating points that are very distant from neighboring clusters, to 0, indicating points that do not belong to a cluster. The figures reveal that the order of distance measures from worst to best are Hamming, City Block, Euclidean, Pearson Correlation and Cosine respectively. For instance, the maximum silhouette value in Figure 5.5 for Hamming distance is around 0.5, whereas, the silhouette value of Figure 5.8 for Cosine distance ranges between 0.5 and 0.9. The larger silhouette value of the Cosine distance implies that the clusters are separated from neighboring clusters. Figure 5.10 reveals the mean value of distances for different clusters [6], [5]. It is calculated by finding the average of distance values between points within clusters and their neighboring clusters. The higher the mean value, the better clusters we get. It is worth noting that the information Figure 5.10 provides does not prove much on its own because it does not take into consideration points distribution within clusters.

The results of the distance plots in Figures 5.3-5.9, the distance mean values in Figure 5.10 as well as the entropy calculations all reveal that Cosine and Pearson Correlation form better clusters than Euclidean, City Block and Hamming distance measures. Based on this information, we choose Cosine measures for all four data sets.

5.1.4 Experiments Results

Web sessions in all four data sets were divided into clusters using the *k*-means algorithm and according to the Cosine distance measure. This grouping of Web sessions into meaningful clusters helps increase the Markov model accuracy. Table 5.4 below is an extract from the data set D1 clusters. It unveils how the cluster group pages within a session according to their categories. The table columns represent the existence or non-existence of a page in a category. Numbers represent the weights or the number of pages, in that particular session, that belongs to the category. It is worth noting that each of the most common categories is allocated in a cluster with the rest of the categories spread across the 7 clusters. We derived from this result that the number of clusters *k* is fully dependent on the nature of the data and the features selected. Therefore, it is highly not recommended to identify *k* before analyzing the data and identifying the purpose of grouping data into clusters.

5.1.5 Comparing IMC, Clustering and MM Accuracy

Figure 5.11 compares the Markov model accuracy of the whole data set to Markov model accuracy using clusters based on Euclidean, Correlation and Cosine distance measures with $k = 7$ for data set D1. Figure 5.12, Figure 5.13, Figure 5.14 and Figure 5.15 compare the accuracy of clustering with that of PMM and the integration of Markov model and clustering (IMC) for the four data sets

using Cosine distance measures for the clusters and based on the 2nd order Markov model. The figures demonstrate a decrease in prediction accuracy using clustering alone. This is due in part to the distance measure used and also to non-categorization of Web sessions [9], [10]. The figures also reveal the improvement in IMC precision results over PMM and clustering. Data sets D3 and D4 show more significant accuracy increase between clustering and Markov model based prediction than data sets D1 and D4. Data sets D1 and D4 reveal more conformity in accuracy increase from clustering to PMM, then IMC.

5.1.6 Comparing IMC to a Higher order Markov Model

5.1.6.1 Comparing State space complexity

Section 5.1.5 experiments prove that the IMC Integration model improves the accuracy of the lower order Markov Model. In this section we experiment further to prove that the IMC Integration model improves the state space complexity of a higher order Markov model. Table 5.5 compares IMC state space complexity.

5.1.6.2 Comparing Accuracy

Acknowledging the fact that IMC improves the prediction accuracy of a lower order Markov model draws our attention to whether or not IMC provides better accuracy than a higher order Markov model. Figure 5.15 reveals the prediction accuracy of IMC as opposed to frequency pruned 3*rd*-order Markov model.

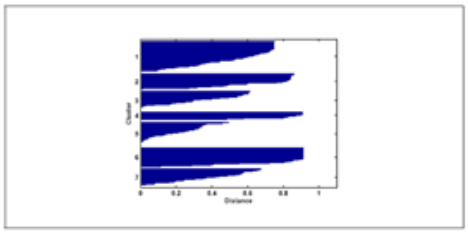

Figure 5.1 Silhouette value of D1 with 7 clusters

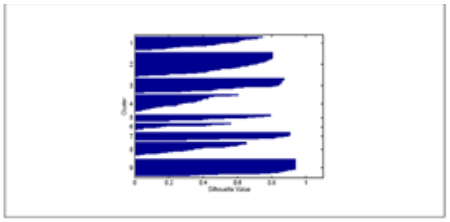

Figure 5.2 Silhouette value of D2 with 9 clusters

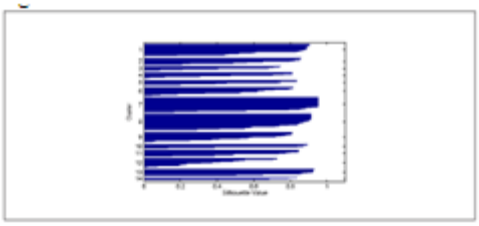

Figure 5.3 Silhouette value of D3 with 14 clusters

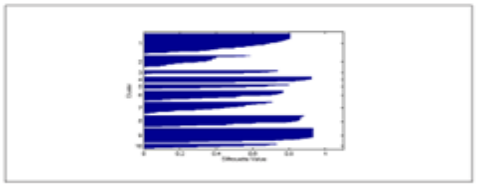

Figure 5.4 Silhouette value of D3 with 10 clusters

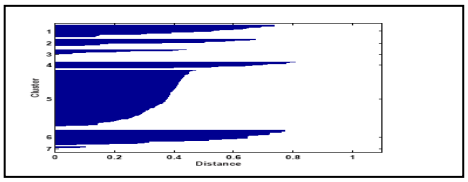

Figure 5.5 Silhouette value of Euclidean distance measure with 7 clusters

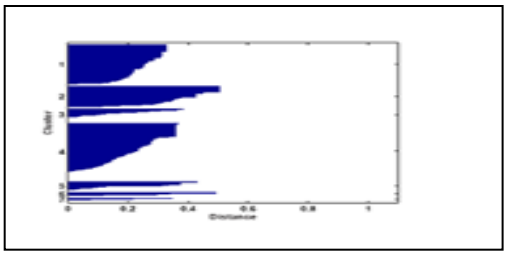

Figure 5.6 Silhouette value of Hamming distance measure with 7 clusters

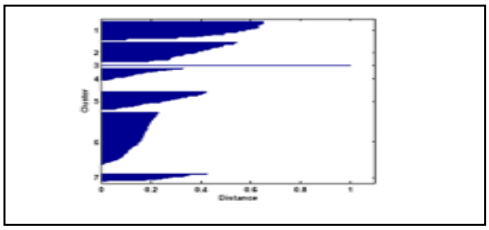

Figure 5.7 Silhouette value of City Block distance measure with 7 clusters

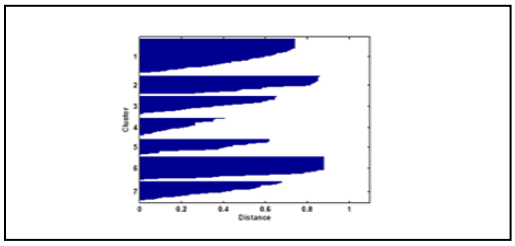

Figure 5.8 Silhouette value of Correlation distance measure with 7 clusters

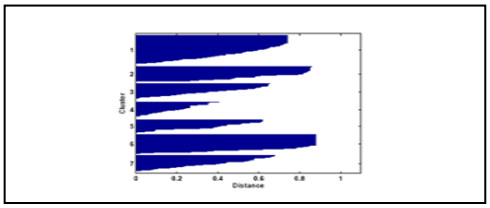

Figure 5.9 Silhouette value of cosine distance measure with 7 clusters

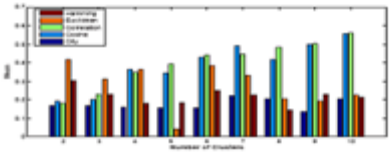

Figure 5.10 the mean value for 2…. 10 clusters using different distance measures

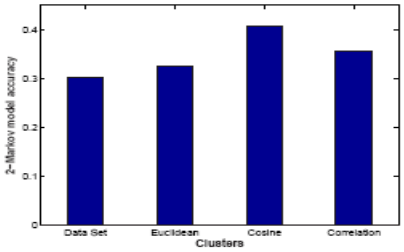

Figure 5.11 Accuracy of clustering, Markov model of whole data set and Markov model accuracy using clusters based on Euclidean, Correlation and Cosine distance measures with k = 7 for data set D1.

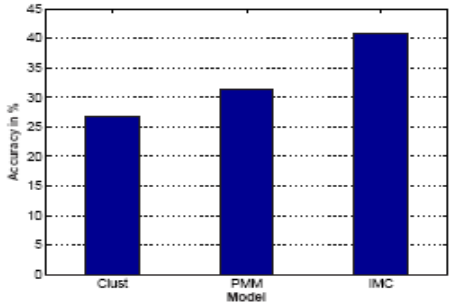

Figure 5.12 Accuracy of clustering, PMM and IMC for data set D1.

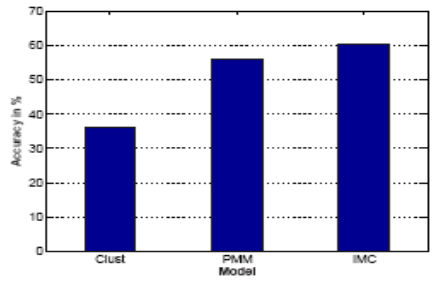

Figure 5.13 Accuracy of clustering, PMM and IMC for data set D3

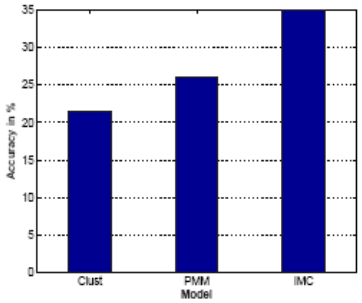

Figure 5.14 Accuracy of clustering, PMM and IMC for data set D4. That of the frequency pruned 3*rd*-order Markov model

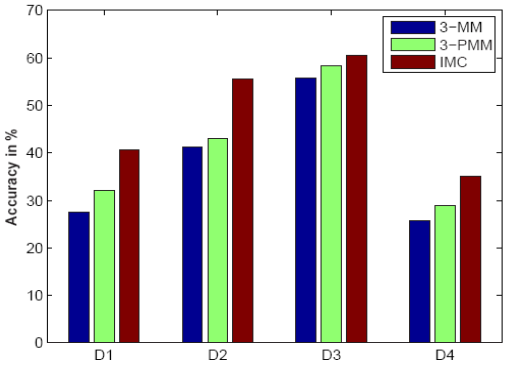

Figure 5.15 Accuracy of 3*rd* order Markov model (3-MM), frequency pruned all 3*rd* order Markov model (3-PMM) and IMC model for all four data sets

Table 5.3 Entropy measures for different clusters

Clusters	2	3	4	5	6	7	8	9	10	20	30	40	50
Euclidean	0.42	0.38	0.32	0.58	0.31	0.28	0.25	0.30	0.26	0.21	0.19	0.23	0.22
City	0.52	0.48	0.50	0.49	0.46	0.42	0.39	0.31	0.29	0.27	0.25	0.24	0.23
Hamming	0.56	0.49	0.53	0.50	0.47	0.39	0.41	0.38	0.36	0.29	0.25	0.31	0.34
Cosine	0.36	0.32	0.37	0.43	0.25	0.21	0.22	0.21	0.17	0.16	0.19	0.22	0.23
Correlation	0.30	0.28	0.30	0.37	0.20	0.21	0.23	0.19	0.20	0.19	0.18	0.19	0.21

Table 5.4 Web sessions grouped into seven clusters

	Access	enviro	EPA	humd	OSW	Press	Waisicons
Cluster 1	-	7	-	-	-	-	-
Cluster 1	-	5	-	-	-	-	-
Cluster 1	-	21	-	-	-	-	-
Cluster 1	-	3	-	-	-	-	-
Cluster 1	-	13	-	-	-	-	-
Cluster 1	-	1	-	-	-	-	-
Cluster 2	-	-	5	-	-	-	-
Cluster 2	-	-	27	-	-	-	-
Cluster 2	-	-	4	-	-	-	-
Cluster 2	-	-	2	-	-	-	-
Cluster 2	-	-	1	-	-	-	-
Cluster 2	-	-	16	-	-	-	-
Cluster 3	-	-	-	-	-	3	-
Cluster 3	-	-	-	-	-	3	-
Cluster 3	-	-	-	-	-	9	-
Cluster 3	-	-	-	-	-	11	-
Cluster 3	-	-	-	-	-	20	-
Cluster 3	-	-	-	-	-	6	-
Cluster 4	-	-	-	-	4	-	-
Cluster 4	-	-	-	-	4	-	-
Cluster 4	-	-	-	-	4	-	-
Cluster 4	-	-	-	-	9	-	-
Cluster 4	-	-	-	-	2	-	-
Cluster 4	-	-	-	-	4	-	-
Cluster 5	-	-	-	4	-	-	-
Cluster 5	-	-	-	5	-	-	-
Cluster 5	-	-	-	11	-	-	-
Cluster 5	-	-	-	6	-	-	-
Cluster 5	-	-	-	9	-	-	-
Cluster 5	-	-	-	4	-	-	-
Cluster 6	-	-	-	-	-	-	3
Cluster 6	-	-	-	-	-	-	12
Cluster 6	-	-	-	-	-	-	3
Cluster 6	-	-	-	-	-	-	1
Cluster 6	-	-	-	-	-	-	4
Cluster 6	-	-	-	-	-	-	2
Cluster 7	8	-	-	-	-	-	-
Cluster 7	11	-	-	-	-	-	-
Cluster 7	12	-	-	-	-	-	-
Cluster 7	8	-	-	-	-	-	-
Cluster 7	3	-	-	-	-	-	-
Cluster 7	4	-	-	-	-	-	-

Table 5.5 IMC number of states

	D1	D2	D3	D4
3-PMM	14,977	18,121	11,218	19,032
IMC	11,682	10,388	19,035	13,634
3-MM	72,524	89,815	50,971	90,123

6. Conclusion

This paper has presented our improvement of markov model accuracy by grouping Web sessions into clusters .The web pages in the user sessions are first allocated into categories according to web services that are functionally meaning full. Then k-means clustering algorithm is implemented using the most appropriate number of clusters and distance measure. Markov model techniques are applied to each cluster as well as to the whole data set. The experimental results reveal that implementing the k-means clustering algorithm on the data sets improves the accuracy of a lower order markov model while reducing the state space complexity of a higher order markov model. The prediction accuracy achieved is an improvement to the previous research papers that addressed mainly recall and coverage.

References

[1] Silky Makker and R K Rathy. Article:" Web Server Performance Optimization using Prediction Prefetching Engine." International Journal of Computer Applications 23(9):19–24, June 2011. Published by Foundation of Computer Science. BibTeX

[2] The Next Page Access Prediction Using Makov Model "International Journal of Electronics Communication and Computer Technology (IJECCT) Volume 1 Issue 1 | September 2011 ISSN: 2249-7838

[3] Agrawal, R. & Srikant, R. (1994), 'Fast algorithms for mining association rules', VLDB'94, Chile pp. 487–

[4] Agrawal, R. & Srikant, R. (1996), 'Mining sequential patterns', International Conference on Data Engineering (ICDE), Taiwan

[5] Albanese, M., Picariello, A., Sansone, C. & Sansone, L. (2004), 'Web personalization based on static information and dynamic user behavior', WIDM'04, USA pp. 80–87.

[6] Ball, G. H. & Hall, D. J. (1965), 'Isodata, a novel method of data analysis and classification', Tech. Rep., Stanford University, Stanford, CA. .

[7] Banerjee, A. & Ghosh, J. (2001), 'Clickstream clustering using weighted longest common subsequences', SIAM Conference on Data Mining, Chicago pp. 33– 40.

[8] Eirinaki, M., Lampos, C., Paulakis, S. & Vazirgiannis, M. (2004), 'Web personalization integrating content semantics and navigational patterns', WIDM'04 pp. 2–9.

[9] Eirinaki, M., Vazirgiannis, M. & Kapogiannis, D. (2005), 'Web path recommendations based on page ranking and markov models', WIDM'05 pp. 2–9.

[10] Gunduz, S. & OZsu, M. T. (2003), 'A web page prediction model based on clickstream tree representation of user behavior', SIGKDD'03, USA pp. 535–540.

[11] Halkidi, M., Nguyen, B., Varlamis, I. & Vazirgiannis, M. (2003), 'Thesus: Organizing web document collections based on link semantics', The VLDB Journal 2003(12), 320–332.

[12] Strehl, A., Ghosh, J. & Mooney, R. J. (2000), 'Impact of similarity measures on web-page clustering', AI for Web Search pp. 58–64.

[13] Wang, K. & Liu, H. (1997), 'Schema discovery for semi-structured data', KDD'97 pp. 271–274.

[14] Wang, K. & Liu, H. (1998), 'Discovering typical structures of documents: A road map approach', SIGR'98 pp. 146–154.

[15] Wang, Q., Makaroff, D. J. & Edwards, H. K. (2004), 'Characterizing customer groups for an e-commerce website', EC'04, USA pp. 218–227.

[16] Zhu, J., Hong, J. & Hughes, J. G. (2002b), 'Using markov models for web site link prediction', HT'02, USA pp. 169–170.

[17] Zuckerman, I., Albrecht, D. & Nicholson, A. (1999), 'Predicting users' request on the www', International Conference on User Modeling (UM99) pp. 275–284.

[18] Adami, G., Avesani, P. & Sona, D. (2003), 'Clustering documents in a web directory', WIDM'03, USA pp. 66–73.

[19] Agrawal, R., Imielinski, T. & Swami, A. (1993), 'Mining association rules between sets of items in large databases', ACM SIGMOD Conference on Management of data pp. 207–216.