Integrating Web Content Clustering into Web Log Association Rule Mining*

Jiayun Guo, Vlado Kešelj, and Qigang Gao

Faculty of Computer Science, Dalhousie University,
6050 University Avenue, Halifax, NS, Canada B3H 1W5
{jguo,vlado,qggao}@cs.dal.ca,
WWW home page: http://www.cs.dal.ca/~{jguo,vlado,qggao}

Abstract. One of the effects of the general Internet growth is an immense number of user accesses to WWW resources. These accesses are recorded in the web server log files, which are a rich data resource for finding useful patterns and rules of user browsing behavior, and they caused the rise of technologies for Web usage mining. Current Web usage mining applications rely exclusively on the web server log files. The main hypothesis discussed in this paper is that Web content analysis can be used to improve Web usage mining results. We propose a system that integrates Web page clustering into log file association mining and uses the cluster labels as Web page content indicators. It is demonstrated that novel and interesting association rules can be mined from the combined data source. The rules can be used further in various applications, including Web user profiling and Web site construction. We experiment with several approaches to content clustering, relying on keyword and character n-gram based clustering with different distance measures and parameter settings. Evaluation shows that character n-gram based clustering performs better than word-based clustering in terms of an internal quality measure (about 3 times better). On the other hand, word-based cluster profiles are easier to manually summarize. Furthermore, it is demonstrated that high-quality rules are extracted from the combined dataset.

1 Introduction

Web Mining is an important application of data mining in the web environment. The problems in this research area became very important due to the immense size of the web resources and intensive user activity. The general area of Web Mining is typically divided into the sub-areas of:

- Web Content Mining, which is concerned with the content of Web pages,
- Web Stucture Mining, concerned with the link structure of the Web, and
- Web Usage Mining, concerned with the patterns of user behaviour when using the Web.

* This work is supported by NSERC.

The data source for Web Usage Mining are typically web server log files. For each user access, a log file includes information such as the user IP number, time of access, file path, user browsing agent, returned status, and the size of transferred data. Data mining on this data set can discover frequent patters in user access, but they are mostly content oblivious. The file path may be a content indicator, but it may not be reliable. Although many Web sites are constructed according to their content, there are also many others that are not. As shown in Table 1,[1] the directory structure of a web site may be organized in different ways. Web site organization can be based on content such as product

Table 1. Different Approaches to Web Site Organization

	Organized by	Server location	Examples
(1)	Product	www.microsoft.com	/windows, /games, /sql...
(2)	Location	www.ibm.com	/us, /ca/en, /cn, /jp...
(3)	Person	www.cs.dal.ca	/~prof01, /~stydent03
(4)	Other	forums.devshed.com	/showthread.php?p-2119#post211...

information ((1) in Table 1), business locations (2), user space (3), and many other conceptual hierarchies. If a Web site is organized according to user space, different directories may contain similar content; e.g., same products or services provided at different places, professors teaching the same class or sharing similar research interests. Also, with the development of Web design techniques, more and more CGI programs are used instead of the traditional static HTML files. In this case ((4) in Table 1), Web pages are generated according to a set of input parameters. Typical examples would be BBS/forum systems. In this case, it may be hard to make any inference about page content based on the URL path.

In this paper we propose and evaluate an approach to integrate the content of Web pages into mining of the Web server log files. We experiment with two different clustering approaches to conceptually organize web pages, and then use cluster labels in association rule mining from the Web log file. The cluster labels represent content information, which is merged with the log file, giving an integrated log file.

2 Related Work

Web Mining is categorized into three categories according to what part of the web is mined [1, 2]: Web Content Mining [3], which focuses on the discovery of useful information from the Web contents; Web Structure Mining [4], which attempts to discover the model underlying the link structure of the Web; and

[1] For privacy reasons, the user information from the www.cs.dal.ca data is anonymized.

Web Usage Mining [5], which attempts to discover knowledge from the data generated by the Web surfer's activity. Web site servers generate a large volume of data from user accesses, which is used to mine knowledge about user browsing behaviour.

Web usage mining is still relatively isolated area from other two areas of Web mining, even though it seems obvious that it is intrinsically related to the page content. For example, the knowledge about user profiles is considered to be a part of Web usage mining [3] and it is hard to learn something useful about user profiles without consulting the content of the visited pages. In analyzing user interaction and profile data, Web usage mining uses only the URL links in the log file, for instance, as indication of the Web page contents. Some ideas of integrating Web content information into Web usage mining have been expressed in some papers like [6–8]. However, most of these attempts still do not use much of information from really looking at the Web pages contents. They either assume that the URLs strongly indicate the Web page contents [7], or use information from log files, like user click streams, to build Web page models or clusters [8]. There are also other attempts to improve the Web server log file mining by integrating some semantic concepts [9, 10], which requires awareness of the content of Web pages beforehand.

3 System Design

Figure 1 illustrates the overall design of our system.

3.1 Preprocessing

In the preprocessing step, there are two major tasks: re-formatting the log file and retrieving Web pages to a local disk-space. Log file re-formatting involves revising the log file to a suitable format for further steps. Each field in the log access log file is revised in order to reduce the cardinality of the corresponding domain set. E.g., the IP numbers are generalized to their sub-net mask consisting of the first two numbers, the access time is discretized into a 4-valued set {morning, afternoon, evening, night}, the dates are grouped into the seven days of week bins, and the numerosity of file paths is reduced by using only their prefix sub-paths. Web page retrieving involves reading the URL address parts of the log file, retrieving the associated Web pages, and storing them to the local space. Only hypertext and plain text files are considered in this phase.

3.2 Document Clustering

Before document clustering, several steps are performed including eliminating HTML/XML tags, eliminating stop words and word stemming for word-based clustering, and translating all letters into their lowercase version for character n-gram-based clustering.

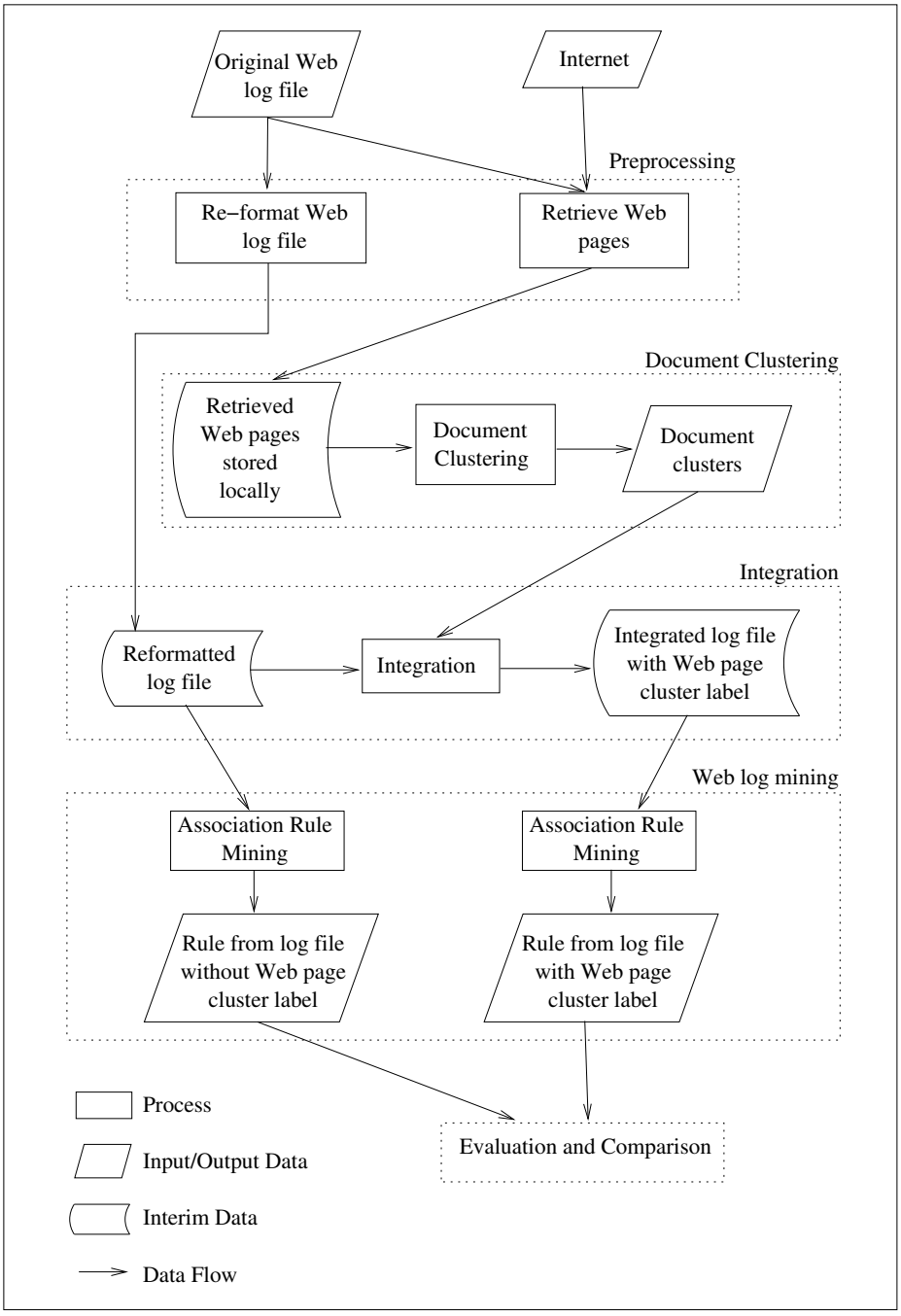

Fig. 1. System Architecture

Document clustering requires vector representation of the documents. For vector components, we use the standard TF-IDF measure, defined in the following way [11]:

$$TFIDF(i,j) = tf(i,j) \cdot \left(1 + \log \frac{N}{df(j)}\right)$$

where $tf(i,j)$ is the frequency of feature (term) t_j in document d_i, N is the number of documents in the collection, and $df(j)$ is document frequency, i.e., the number of documents in the collection containing the term t_j.

In the vector space model, one of the most common measures for similarity between documents is the cosine measure, defined by [12]

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{||d_1|| \cdot ||d_2||}$$

where d_1 and d_2 are two document vectors. The measure returns values close to 0 for very dissimilar documents, and high values close to 1 for very similar documents. The k-means clustering algorithm requires a distance measure which produces distance close to 0 for very similar documents and higher values for dissimilar documents, we use the sinus distance measure [12]:

$$\sin(d_1, d_2) = \sqrt{1 - \cos(d_1, d_2)^2}$$

Another obvious option would be simply to use $1 - \cos(d_1, d_2)$, but a more detailed analysis shows that the sinus measure is more appropriate since the k-means algorithm relies on centroid calculation, which is a linear transformation, and since $\sin(x)/x \to 1$ when $x \to 0$.[2]

The well-known K-means algorithm is used for document clustering and its pseudo code is given in Algorithm 1. A centroid is calculated as the arithmetic mean and mass-center of all points in a cluster. There are typically three options for a stopping criterion: We may stop when clusters settle. Since clusters may oscillate instead of settling at a fixed point, there is a limit on the maximal number of iterations. The third option is to observe a clustering quality measure and stop when this measure reaches a local maximum. We use the fixed point criterion with a limit on the number of iterations.

Algorithm 1 K-means

Partition object into k non-empty subsets randomly
repeat
  Compute the centroids of the clusters
  Assign each object to the cluster with the nearest centroid
until some stop criterion is met

[2] Another way of presenting the argument is to depict document clustering as clustering of points at the surface of a unit n-dimensional sphere.

The clustering quality is evaluated using either *external* quality measures, which rely on some external knowledge such as a gold clustering standard of a data set, or *internal* quality measures, which do not rely on any external knowledge. Since our set of web pages is not labeled, we first use an internal evaluation function to evaluate clustering. After using clustering results in the web log association rule mining, we evaluate the final results manually, and this is an external evaluation. However, unlike the internal evaluation, it is a qualitative rather than quantitative evaluation.

For the internal evaluation, we use the internal evaluation function proposed in [13] and defined as

$$EI = \frac{1}{mN} \sum_{j=1}^{N} E(t_j) \sum_{i=1}^{m} E_i(t_j) \tag{1}$$

where $E(t_j)$ and $E_i(t_j)$ are defined in the following way:

$$E(t_j) = 1 - \frac{m_j - 1}{m - 1} \log_m \sum_{i=1}^{m} E_i(t_j)$$

$$E_i(t_j) = \frac{n_{ij} - 1}{n_i - 1} \log_{f_{i\ max}+1}(\overline{f_j} + 1)$$

In the equations above, N is the number of documents, m is the number of clusters, $E(t_j)$ is called inter-cluster entropy, $E_i(t_j)$ is called intra-cluster entropy, n_{ij} is the number of documents including feature t_j in cluster C_i, $f_{i\ max}$ is the maximum frequency of feature t_j in cluster C_i, $\overline{f_j}$ is the average frequency of feature t_j in cluster C_i, and m_j is the number of clusters in which feature t_j appears.

3.3 Integration

The integration step involves integration of the Web document cluster information into log files. Two data sets are obtained for further mining: one is *log_origin* with only the information obtained from the web log file, and the other is named *log_integ* and it includes information from the web log file, integrated with the cluster labels. In order to be able to interpret the results of the association rule mining, the clusters are manually summarized and described by brief descriptive paragraphs in plain language.

3.4 Association Rule Mining

In this last step, the Apriori association rule mining algorithm [14] is applied to the two data sets obtained from the above steps.

The number of unique values of each field in the re-formatted log file is limited, and all the values can be displayed as strings. They are then used as *items* in the standard association rule mining terminology. Each different value

either is present or not, so it is treated as a Boolean value. Since the domain sets of different fields are disjoint, we do not need to present field (attribute) names when presenting the association rules. The purpose of this association mining step is to discover the rules of co-occurrence and the implications underlying the large amount of access records.

After applying association rule mining on the two datasets *log_origin* and *log_integ*, two sets of rules were obtained from the datasets respectively. These two sets of rules are compared. The rules obtained from the dataset *log_origin* are a subset of the rules obtained from the dataset *log_integ*, so we explore the rules obtained from *log_integ* but not from *log_origin* to see whether they provide any useful and novel information.

4 Results and Evaluation

We used an Apache log access file from the graduate Web server of the Faculty of Computer Science at Dalhousie University, for a one-month period in October 2003. In this period, there were 161,499 access records producing a 230MB log file. Using a widely used data set in the experiment would be beneficial for comparative reasons with other published results, however we were not able to locate such dataset that would involve both web pages and the web log data. In other published work, this scenario is often seen, were the experiments are based on a local departmental web server.

All the experiments are executed on a Sun Solaris server at the CS Faculty of Dalhousie University. The server type is SunOS sparc SUNW, Sun-Fire-880. The system was implemented using Perl (preprocessing, document clustering, and integration) and C++ (Association Rule Mining).

4.1 Cluster Summarization

In the document clustering step, after the K-means algorithm is performed, the frequencies of features in each cluster are obtained. From these, the most frequent key features of each cluster are extracted and used to manually summarize the major topic of each cluster. In the manual summarization, beside the set of the most frequent features, some sample Web pages from a cluster are examined in order to produce a reliable cluster summary. Even though we could not successfully use any existing summarization tool, a part of future work is to make a further attempt to use this option.

The analysis of these summaries produced to following observations:

(1) When the number of clusters k is relatively large (k=12 for k-means) with word representation, some different partitions share same or similar topics.

(2) When k is relatively small (k=8 for k-means) with word representation, some of important clusters, which appear when k is larger, were not partitioned from the others.

(3) The optimal clustering summaries are obtained for $k = 10$ with the word representation.

(4) When the character n-gram representation is used, it is very difficult to summarize clusters manually.

4.2 Internal Cluster Evaluation

The results of the internal cluster evaluation using equation 1 are shown in Table 2. We can see that k=10 produced the best results for both word and n-gram representation, which is an interesting result since it coincides with our analysis based on cluster summaries. Character n-gram representation produced significantly better results (3 times) than word representation. However, since it is much harder to summarize character n-gram based clusters than word-based clusters, so we chose to proceed with the word clusters to the association rule mining step. An important open question is how to summarize clusters based on their character n-gram profiles. If this problem could be successfully solved our hope is that we would obtain even better association rules.

Table 2. Comparison of Document Clustering

	Word-rep	Ngram-rep
K=8	0.00953	0.03478
K=10	0.01183	0.036515
K=12	0.01077	0.03556
K=14	0.01032	0.03427

4.3 Association Rule Mining Evaluation

In the association rule mining step, after applying Apriori on both datasets of *log_origin* and *log_integ*, we got two lists of association rules. Table 3 shows the number of association rules obtained from the two datasets.

Table 3. Number of rules obtained

[support, confidence]	Log_origin	Log_integ
[2%, 30%]	64	203
[2%, 50%]	20	81
[1%, 50%]	37	187
[1%, 60%]	9	58

The integrated log file produce three to four times more rules than the original log file. As all the attributes in *log_origin* are also included in *log_integ*, it is obvious that the rules from the latter are also included in those from the former. Since the number of access records is very large, we mine rule with a support threshold of only 1 or 2%, however the confidence threshold is kept at higher levels of 30, 50, and 60%.

Table 4 and Table 5 list some rules obtained from the two datasets, *log_origin* and *log_integ* respectively. The left side columns display the rules obtained from program, while the right side columns display the same rules interpreted in the plain language. Since all the rules obtained from *log_origin* are also included in the rules obtained from *log_integ*, in Table 5 we show only the rules that are not obtained from *log_origin*.

We can make interesting observations about the web site usage based on the extracted rules. According to Table 5, the rules indicate when and from where the access queries occurred, who visited, and what kind of information was requested. These rules provide information which can be used in various applications, including web site organization, Web content distribution, and analysis of user access behavior. For example, from the rule "/~prof33 ⇒ ERROR", we would conclude that *prof33* had changed a lot of his web pages, and we may suggest an update or creation of redirection Web pages under his domain. The rules like "cluster5 ⇒ /~prof11", "cluster6 ⇒ /~prof07", provided information about the user domains that provide the content of a certain category or certain topic. The rules such as "24.222 ⇒ cluster7" and "156.34 ⇒ cluster7" tell us about the topics of interest of visitors from certain internet domains. These rules are related to the document cluster labels, i.e., the web contents, and were not included in the results from the conventional data provided in the web log file (*log_origin*).

Table 4. Association Rules from *log_origin*

K=10 with Word Representation Support=1% Confidence=50%	
Association Rules	Rules in plain language
/~prof12 ⇒ 129.173 [10, 53]	A majority of accesses to user *prof12*'s web pages are from the CS building log-ons
129.173 ⇒ afternoon [20, 51]	Over half of the accesses from CS building were in the afternoon;
/~prof12 ∧ Tue ⇒ afternoon [2, 56] /~prof12 ∧ Wed ⇒ afternoon [2, 52]	Accesses on Tue and Wed to user *prof12*'s web pages occurred mainly in the afternoon;

From the original dataset, we obtained the rules that typically describe certain visitor groups that are interested in certain professors' web pages. However,

Table 5. Association Rules from *log_integ*

K=10 with Word Representation Support=1% Confidence=50%	
Association Rules	Rules in plain language
cluster5 ⇒ /~prof11 [3, 51]	A majority of Java programming pages are from user *prof11*.
/~prof11 ⇒ cluster7 [6, 54]	A majority of user *prof11*'s web pages are personal or course information pages.
cluster6 ⇒ /~prof07 [3, 86] /~prof07 ⇒ cluster 7 [10, 59]	User *prof07* provides over eighty percent of administration pages, however more than half of user *prof07*'s web pages are personal or course information pages.
/~prof33 ⇒ ERROR [4,81]	User *prof33* has deleted or modified many of his web pages since Oct. 2003.
/~prof13 ⇒ cluster1 [5,78] cluster1 ⇒ /~prof13 [5,67]	User *prof13* has many empty pages.
129.173 ⇒ cluster7 [21,51] 142.177 ⇒ cluster7 [3,52] 156.34 ⇒ cluster7 [2,55] 24.138 ⇒ cluster7 [1,56] 24.215 ⇒ cluster7 [1,54] 24.222 ⇒ cluster7 [7,55] 24.224 ⇒ cluster7 [3,53]	Over 50% of accesses from outside CS building are for general information.
/~prof10 ∧ afternoon ⇒ cluster7 [1,53]	A majority of accesses to user *prof10*'s web pages in afternoon are for general information.
/~prof12 ∧ cluster0 ⇒ 129.173 [1,60] /~prof12 ∧ 24.222 ⇒ cluster7 [1,62]	

one single professor's web site may contain different topics. From the integrated dataset, we obtained the rules that contain information about visitor groups that are interested in certain kinds of topics. Pages with similar topics may exist in different professors' directories, and these rules are not found from the original dataset.

Therefore, we can conclude that we demonstrated that some useful rules are obtained from integrating web document clusters and web log files. These rules are related to the content of web pages, and provide information that can be further used for user profiling and web site evaluation and improvement.

5 Conclusion and Future Work

In this paper, a novel approach to Web log file mining combined with the information from automatic Web page clustering is presented. The methods for document clustering are used: word-based and character n-gram based. The K-means algorithm was used in web page clustering. After manually summarizing clusters obtained form the web log file, and from the integrated data file, the Apriori association rule mining algorithm is applied. Several evaluation results are produced: an "optimal" number of clusters is found based on manual summarization and cluster analysis, and it was confirmed that this number of clusters is locally optimal in terms of the internal quality measure. Furthermore, it was demonstrated that some interesting content-related rules can be discovered from the integrated web log data, while they could not be discovered using only the standard web log data. These rules provide useful information related to the web usage mining, and can be useful in tasks of the web site organization, web content distribution, customer behaviour profile, and similar.

The designed system is a proof-of-a-concept prototype of the idea of combining the web content mining and web usage mining, and there are many obvious aspects in which it can be improved:

- The algorithm should be improved to handle larger data sets.
- More types of files should be analyzed, beside HTML and plain text only.
- Automatic summarization technique should be applied.
- Generating summaries for n-gram based clusters would open the doors of using better clustering results in rule mining.
- The use of concept hierarchies could improve quality or association rule mining.
- The data mining functionalities other than association rule mining could be used in web log analysis.

Acknowledgments

We would like to thank Haibin Liu and anonymous reviewers for providing useful comments. The authors gratefully acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Madria, S., Bhowmick, S., Ng, W., Lim, E.: Research issues in web data mining. In: Proceedings of Data Warehousing and Knowledge Discovery, First International conference, DaWaK'99. (1999) 303–312
2. Borges, J., Levene, M.: Data mining of user navigation patterns. In: Proc. of WEBKDD'99 ws. on Web Usage Analysis and User Profiling. (1999) 92–111
3. R.Kosala, H.Blockeel: Web mining research: A survey. ACM SIGKDD **2** (2000) 1–15
4. Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S., Raghavan, P., Rajagopalan, S., Tomkins, A.: Mining the link structure of the World Wide Webx. IEEE Computer **32** (1999) 60–67
5. Cooley, R., Mobasher, B., Srivastava, J.: Web mining: Information and pattern discovery on the world wide web. In: Proc. of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97). (1997) 558–567
6. Mobasher, B., Dai, H., Luo, T., Sun, Y., J.Zhu: Integrating web usage and content mining for more effective personalization. In: Proc. of the Intl. Conf. on Ecommerce and Web Technologies (ECWeb). (2000) 165–176
7. Kato, H., Nakayama, T., Yamane, Y.: Navigation analysis tool based on the correlation between contents distribution and access patterns. In: Proc. of the Web Mining Workshop KDD00. (2000) 95–104
8. Ypma, A., Heskes, T.: Categorization of web pages and user clustering with mixtures of hidden markov models. In: Workshop on Web Knowledge Discovery and Data mining (WEBKDD 2002). (2002) 31–43
9. Jin, X., Zhou, Y., Mobasher, B.: A unified approach to personalization based on probabilistic latent semantic models of web usage and content. In: Proc. of the AAAI 2004 Workshop SWP'04. (2004) pp. 26–34
10. Eirinaki, M., Lampos, C., Paulakis, S., Vazirgiannis, M.: Web personalization integrating content, semantics and navigational patterns. In: ACM Web Information and Data Management Workshop. (2004) 72–79
11. Aslton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management **24** (1988) 513–523
12. Miao, Y., Keselj, V., Milios, E.: Comparing document clustering using n-grams, terms and words (2004)
13. Jo, T.: Evaluation function of document clustering based on term entropy. In: Proc. of 2nd International Symposium on Advanced Intelligent System. (2001) 95–100
14. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers (2001)
15. M.Steinbach, G.Karypis, V.Kumar: A comparison of document clustering techniques. In: Proc. of the Text Mining Workshop, KDD00. (2000)
16. Pandey, A., Srivastava, J., Shekhar, S.: A web proxy server with an intelligent prefetcher for dynamic pages using association rules. Technical Report TR-01-004, Department of Computer Science, University of Minnesota (2001)
17. Porter, M.: An algorithm for suffix stripping. Program **14** (1980) 130–137
18. Etzioni, O.: The World Wide Web: Quagmire or gold mine. Communications of the ACM **39** (1996) 65–68
19. Saltonandand, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM **18** (1975) 613–620
20. Punin, J., Krishnamoorthy, M., M.J.Zaki: Mining web log data across all customers touch points. In: Web Usage Mining—Languages and Algorithms, WEBKDD01 Workshop. (2001) 88–112