WEB USAGE MINING USING ROUGH AGGLOMERATIVE CLUSTERING

Pradeep kumar, P. Radha Krishna
Institute for Development and Research in Banking Technology, (IDRBT),
1, Castle hills, Masab Tank, Hyderabad - 500057

Supriya kumar De
XLRI Jamshedpur, C.H.Area(E),
Jamshedpur,
INDIA

S Bapi Raju
University ofHyderabad,
Gochibowli,
Hyderabad,INDIA

Keywords: Data mining, rough sets, clickstream, web usage mining , similarity upper approximation.

Abstract: Tremendous growth of the web world incorporates application of data mining techniques to the web logs. Data Mining and World Wide Web encompasses an important and active area of research. Web log mining is analysis of web log files with web pages sequences. Web mining is broadly classified as web content mining, web usage mining and web structure mining. Web usage mining is a techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. This paper demonstrates a rough set based upper similarity approximation method to cluster the web usage pattern. Results were presented using clickstream data to illustrate our technique.

1 INTRODUCTION

World Wide Web (WWW) is an unstructured collection of pages and hyperlinks. People from different backgrounds and interests access and provide web pages. Application of data mining approaches on World Wide Web is referred as web mining. Web mining has attracted a lot of researchers due to huge amount of active data available on the World Wide Web. Broadly, web mining tasks include web usage mining, web content mining and web structure mining.

Web content mining is a process of discovering information from millions of sources across the World Wide Web. User interaction on the web are recorded on a web logs. As each user interaction corresponds to a mouse click it is oftenly referred as clickstream. Web usage mining is performing mining on web usage data or web logs. Extracting patterns from on line information, such as HTML files or E-mails is referred as web content mining. Web content mining goes beyond basic Information retrieval technology. Web structure mining is a research field focused on using the analysis of the link structure of the web, and one of its purposes is to identify more preferable documents. The intuition is that a hyperlink from document A to document B implies that the author of document A thinks document B contains worthwhile information.

Like conventional data mining clustering, association and sequential analysis are three important operations in web mining. This paper focuses on clustering, which is a unsupervised learning method to partition a set of patterns into

kumar P., Radha Krishna P., kumar De S. and Bapi Raju S. (2005).
WEB USAGE MINING USING ROUGH AGGLOMERATIVE CLUSTERING.
In *Proceedings of the Seventh International Conference on Enterprise Information Systems*, pages 315-320
DOI: 10.5220/0002553003150320

groups (Bezdek, J., 1981). To show the viability of our approach we applied upper similarity approximation to cluster clickstream transactions.

In this paper, we present an agglomerative clustering approach using upper similarity approximation for mining clickstream data. Clickstream is a sequence of URLs browsed by a user within a particular website in one session. To discover the pattern of groups of users with similar interest and motivation for visiting that particular website can be found by clustering users' clickstream on a particular website. A user session is the clickstream of page views for a single user in the website. We considered each user session as a clickstream transaction, which contains the sequence of URLs (or hyperlinks) of a visitor visiting a web site.

A lot of research has been done in the area of Web Usage Mining (Cooley, R., 2000, Spiliopoulou, 1999, Manco, G et.al., 2003) which directly or indirectly addresses the issues involved in the extraction of web navigational patterns (Spiliopoulou, M. and Faulstich, L. C., 1999), ordering relationships (Mannila, H. and Meek, C., 2000), prediction of web surfing behavior (Pitkow, J and Pirolli, P., 1999), and clustering of web usage sessions (Fu . et. al , 2000) based on web logs, possibly supplemented by web content or structure information. Perkowitz and Etzioni (Perkowitz and Etzioni, 2000) proposed the idea of optimizing the structure of web sites based on co-occurrence patterns of pages within usage data for the site. Spiliopoulou and Cooley (Spiliopoulou, 1999; Cooley, R., 2000) have applied data mining techniques to extract usage patterns from web logs, for the purpose of deriving market intelligence. Well-developed mining techniques cannot be applied directly for web data as web logs being unstructured in nature. Clustering in web mining faces several additional challenges (Jhoshi, A. and Krishnapuram , R., 1998).The specific problem of web usage clustering has been studied over the past few years. In (De and Radha Krishna, 2002), automatic personalization of a web site from user transactions using fuzzy proximity relations is presented. In (De and Radha Krishna, 2004), a clustering algorithm is presented using rough approximation to cluster web transactions from web access logs. Web clusters tends to have fuzzy boundaries. It is likelihood that an object may be a candidate for more than one clusters. To deal with the special challenges found in web usage data a non-conventional clustering approach using rough set theory has been presented in (Hogo, M et al. ,2004). Pawan Lingras (Lingras, P., 2003) has used rough set theory for web mining clustering.

The rest of the paper is organized as follows: section 2 describes the basics of rough set theory. In section 3, we present an approach for grouping clickstream using upper similarity approximation. Experimental results are presented in section 4 and we conclude in section 5.

2 ROUGH SET THEORY

Zdzisław Pawlak introduced Rough set theory (Pawlak ,1982) to deal with uncertainty and vagueness. Rough set theory became popular among scientists around the world due to its fundamental importance in the field of artificial intelligence and cognitive sciences. This section provides a brief summary of the concepts of rough set theory. The building block of rough set theory is an assumption that with every set of the universe of discourse we associate some information in the form of data and knowledge.

Let U denote a universe and let $R \subseteq U \times U$ be a equivalence relation on U. The pair $A = (U,R)$ is called an approximation space. The equivalence relation R partitions the set U into disjoint subsets. Such a partition of the universe is denoted by $U/R = (E_1,E_2,E_3,\ldots,E_n)$, where E_i is an equivalence class of R.. If two elements $u, v \in U$ belong to the same equivalence class $E \subseteq U/R$, we say u,v are indistinguishable. The equivalence classes of R are called the elementary or atomic sets in the approximation space $A = (U,R)$.

Within the same equivalence class it is not possible to differentiate the elements. Hence, one may not get a precise representation for an arbitrary set $X \subseteq U$ in terms of elementary sets in A. Rather its upper and lower bounds may represent the set X. Lower approximation $\underline{A}(X)$ is union of all the elementary sets which are subsets of X.

$$\underline{A}(X) = \{ x \in U : (x) \subseteq X \}$$

The upper bound $\bar{A}(X)$ is union of all the elementary sets that have a non empty intersection with X.

$$\bar{A}(X) = \{ x \in U : (x) \cap X \neq \phi \}$$

The pair $(\underline{A}(X) , \bar{A}(X))$ is the representation of an ordinary set of X in the approximation space $A = (U, R)$ or simply the rough set of X. Fig 1 illustrate the rough set approximation.

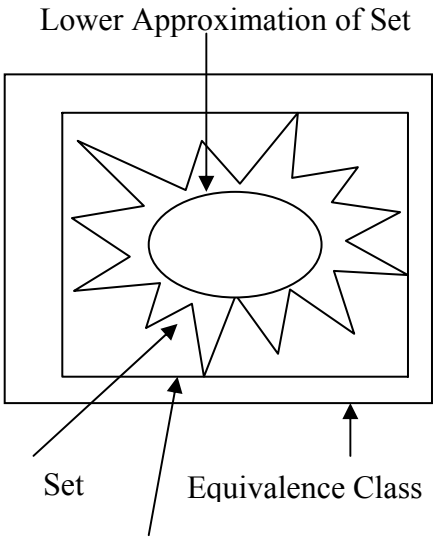

Figure 1: Rough set Approximation

3 CLUSTERING USING ROUGH SETS

In this section, we present the agglomerative clustering for clustering clickstream transactions using upper similarity approximation. In rough set theory, the lower approximation of a concept consists of all objects that definitely belong to the concept. The upper approximations of the concept consist of all objects that possibly belong to the concept. In our approach, we consider upper approximation property to form clusters by defining a similarity upper approximation.

We represent each transaction as a Jaccard vector similarity function. The Jaccard similarity penalizes on a small number of shared clicks. Let t and s be the two clickstream transactions. The similarity between the two transactions is computed as

$$\text{sim}(t,s) = \frac{|t \cap s|}{|t \cup s|}$$

Here, sim(t,s) Є (0,1). sim(t,s) will be equal to 1 when two transaction t and s are exactly same and sim(t,s) is 0 when two transaction t and s are completely dissimilar. The similarity measure provides an idea of interest and motivation of users' access pattern in their common area.

For a given threshold value, $th \in (0,1)$, and for any two user transactions t and $s \in T$, a binary relation R on T denoted as tRs is defined by tRs iff $sim(t, s) \geq th$, where T is a set of all clickstream transactions. The similarity class of t, denoted by R(t), is the set of transaction which are similar to t is given by $R(t) = \{ s \in T, sRt \}$. For a fixed threshold $\in (0, 1)$, a binary tolerance relation R is defined on T.

For clustering clickstream transactions, we compute a similarity upper approximation as follows:

Let $t_i \in T$ be a user clickstream. The upper approximation $\overline{R}(t_i)$ is a set of transactions similar to t_i , that is, a user, who is visiting the hyperlinks in t_i, may also visit the hyperlinks present in other transactions in $\overline{R}(t_i)$. Similarly, $\overline{R}\,\overline{R}(t_i)$ is a set of transactions that are possibly similar to $\overline{R}(t_i)$, and this process continues until two consecutive upper approximations for t_i are same. The process of finding the two equal consecutive upper approximations is known as Similarity Upper Approximation and denoted by S_i.

Initially, each clickstream transaction has been considered as individual cluster. The similarity upper approximation for each clickstream transaction is calculated for a given clickstream transaction data set. In each iteration of agglomerative clustering, the clusters are agglomerates based on the similarity upper approximation. The process of computing similarity upper approximation is repeated for each transaction, until the two consecutive upper approximations are same.

Let $S_1, S_2, S_3, \ldots S_n$ be similarity upper approximation for transaction $t_1, t_2, t_3, \ldots, t_n$ respectively. Now, if $S_i = S_j$ (i and j are distinct) allocate t_i and t_j in the same cluster. Performing this way, we get a distribution of m disjoint clusters. Let these m clusters be C_j (j = 1, 2, …, m). Here, C_j 's are all distinct and $\cup C_j = T$. These C_js represent the sub-groups of the transactions representing the transaction cluster.

Table 1: Similarity Matrix

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
t_1	1	0	0	0.5	0	0	0	0	0	0
t_2	0	1	0	0	0.6	0.4	0	0	0	0
t_3	0	0	1	0	0	0	0	0	0.8	0
t_4	0.5	0	0	1	0	0	0.5	0	0	0
t_5	0	0.6	0	0	1	0	0	0	0	0
t_6	0	0.4	0	0	0	1	0	0	0	0
t_7	0	0	0	0.5	0	0	1	0	0	0
t_8	0	0	0	0	0	0	0	1	0	0.6
t_9	0	0	0.8	0	0	0	0	0	1	0
t_{10}	0	0	0	0	0	0	0	0.6	0	1

The algorithm for clustering clickstream transactions is given below:

Algorithm: Rough Agglomerative Clustering
Input: A set of n objects in a data set $U = \{x_1, x_2, \ldots x_n\}$, Threshold θ, the number of clusters p ($\leq n$)
Output: Cluster scheme C
Step 1 : Start
Step 2 : Initially consider each object of U as a cluster of one member $C_i = \{ x_i\}$ and $C = \{ C_1, C_2, ., C_n \}$
Step 3 : For each pair of clusters C_i and C_j calculate $sim (C_i , C_j) = (C_i \cap C_j) / (C_i \cup C_j)$
Step 4 :For each cluster C_i, find out the similarity upper approximation S_i. for a given threshold θ.
Step 5 : If $S_i = S_j$, form a new cluster $C_{ij} = C_i \cup C_j$, i.e. put x_i and x_j in the same cluster.
Step 6: Update C
Step 7: Repeat Steps 5 and 6 till there is no change in the number of clusters.
Step 8 : Output C
Step 9. Stop

Let N be the total number of clickstream transactions and *L* be the average length of the transaction. The complexity of similarity computation is in the order of $O(N^2 \log_2 L)$. Let R be relation defined over T then the complexity of upper approximation is in the order of O (T/R) (Jamil and Jitender , 2001), which is same as O(N/R). Merging of clusters takes place at each iteration based on the similarity upper approximation. Let k be the average number of clusters merging in each iteration. The complexity of merging k clusters is in the order of O (k log k) (Dash et.al., 2003) and there may be maximum of N/k iterations. Thus, the complexity of merging process is O((N/k) k log k) = O (N log k). So, the complexity of rough agglomerative clustering is of the order $O (N^2 \log_2 L)+ O (N/R) + O (N \log k)$.

To explain the approach, consider navigation patterns of user visiting a e-commerce site shown in transaction set T.

$T = \{ t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}\}$

t_1 ={ Home, Login, Help, Logout }
t_2= {Register, Regport, Results, Regform1, Regform2 }
t_3 ={ Catalog, Product, P_Info, AddCart }
t_4 ={ Home, Login, Help, Fdback, Shelf, Promo, Download, Logout }
t_5 ={Register, Regform1, Results }
t_6 ={Regport, Regform2 }
t_7 ={ Fdback, Shelf, Promo, Download }
t_8 ={ Charge, Pay_req, Pay_rem, Freeze}
t_9 ={Catalog, Product, P_Info, Cart, AddCart }
t_{10} ={Charge, Pay_req, Pay_rem }

Equivalent similarity matrix is shown in Table 1. Computing similarity upper approximation, we get at threshold value 0.4, the equivalence classes are

$R(t_1) = \{ t_1 , t_4 \}$, $R(t_2) = \{t_2 , t_5 , t_6 \}$,

$R(t_3) = \{t_3 , t_9\}$, $R(t_4) = \{t_1, t_4, t_7\}$,

$R(t_5) = \{ t_2 , t_5 \}$, $R(t_6) = \{ t_2 , t_6\}$,

$R(t_7) = \{t_4 , t_7\}$, $R(t_8) = \{t_8 , t_{10}\}$,

$R(t_9) = \{t_3 , t_9\}$, $R(t_{10}) = \{ t_8 , t_{10}\}$.

In the first step, we compute the upper approximation of all ten transactions.

$\overline{R}(t_1) = \{ t_1 , t_4 \}$, $\overline{R}(t_2) = \{t_2 , t_5 , t_6 \}$,

$\overline{R}(t_3) = \{t_3 , t_9\}$, $\overline{R}(t_4) = \{t_1, t_4, t_7\}$,

$\overline{R}(t_5) = \{ t_2 , t_5 \}$, $\overline{R}(t_6) = \{ t_2 , t_6\}$,

$\overline{R}(t_7) = \{t_4 , t_7\}$, $\overline{R}(t_8) = \{t_8 , t_{10}\}$,

$\overline{R}(t_9) = \{t_3 , t_9\}$, $\overline{R}(t_{10}) = \{ t_8 , t_{10}\}$.

Computing similarity upper approximation, we get

$\overline{R}\ \overline{R}(t_1) = \{ t_1, t_4, t_7 \}$, $\overline{R}\ \overline{R}(t_2) = \{t_2 , t_5 , t_6\}$,

$\overline{R}\ \overline{R}(t_3) = \{t_3 , t_9\}$, $\overline{R}\ \overline{R}(t_4) = \{ t_1, t_4, t_7 \}$,

$\overline{R}\ \overline{R}(t_5) = \{t_2 , t_5 , t_6\}$, $\overline{R}\ \overline{R}(t_6) = \{t_2 , t_5 , t_6\}$,

$\overline{R}\ \overline{R}(t_7) = \{ t_1, t_4, t_7 \}$, $\overline{R}\ \overline{R}(t_8) = \{ t_8 , t_{10}\}$.,

$\overline{R}\ \overline{R}(t_9) = \{t_3 , t_9\}$, $\overline{R}\ \overline{R}(t_{10}) = \{ t_8 , t_{10}\}$.

$\overline{R}\ \overline{R}\ \overline{R}(t_1) = \{ t_1, t_4, t_7 \}$,

$\overline{R}\ \overline{R}\ \overline{R}(t_5) = \{t_2 , t_5 , t_6\}$,

$\overline{R}\ \overline{R}\ \overline{R}(t_6) = \{t_2 , t_5 , t_6\}$,

$\overline{R}\ \overline{R}\ \overline{R}(t_7) = \{ t_1, t_4, t_7 \}$,

Now, the process stops as two consecutive upper approximations for each transaction is same. Thus, the clusters formed are $\{ t_1, t_4, t_7 \}$, $\{t_2 , t_5 , t_6\}$, $\{t_3 , t_9\}$, and $\{ t_8 , t_{10}\}$, that is, we have four clusters. Since two or more clusters will agglomerate at each stage the algorithm converges faster. Below we describe the mean profile of each cluster.

Cluster1: It consists of three user navigation pattern t_1 , t_4 , t_7. Although both the t_1 and t_7 has navigated different set of pages but with respect to t_4 both has navigated at least 40% similar pages.

Cluster 2: It consists of three user navigation pattern t_2 , t_5 , t_6. All the three perform the same navigation pattern at least 40% with respect to one another.

Cluster 3: It consists of two user navigation pattern t_3 and t_9. Both have them have navigated product information site and their navigation pattern is at least 40% similar.

Cluster 4: It consists of two user navigation pattern t_8 and t_{10}. Both have them have navigated product information site and their navigation pattern is at least 40% similar.

4 EXPERIMENTAL RESULTS

We implemented our approach using Java and performed experiments on a 2.4 GHz, 256 MB, Pentium-IV machine running on Microsoft Windows XP 2002. We used the clickstream dataset T40I10D100K(http://www.cs.helsinki.fi/u/goethals /dmcourse/util.html.), a Hungarian on-line news portal. The dataset contains 1,00,000 clickstream transactions. This set can be generated using the generator from the IBM Alamaden Quest Research group(http://www.almaden.ibm.com/software/quest/ Resources/index.shtml). The clickstream transaction dataset contains transaction as small as one click and as large as thirty clicks. The average weighted length of the clicks is 10.06. Intuitively, very small and very large clickstreams may not provide any useful information about the users' navigation behavior. Thus, transaction length having less than 5 clicks is considered as a short transaction and transaction length with greater than 15 clicks are considered as a long transaction. In the preprocessing step, short and long transactions are removed from the dataset.

Experiments are performed on preprocessed dataset with 81,832 records. At threshold value 0.8 we got 1,131 clusters and it took around 15hours 58 minutes and 17 seconds. We randomly took 2000 records preprocessed it, at 0.29 thershold value we got 154 clusters. Similarly, we took randomly 50,000 records preprocessed it , at threshold value 0.6 we got 1520 clusters.

5 CONCLUSIONS

Clustering is the task of grouping similar objects into clusters. Hierarchical agglomerative clustering approaches iteratively agglomerates the closest (or similar) pair of clusters. In this work, we presented a rough agglomerative clustering technique to cluster clickstream transactions based on Upper similarity approximation. We experimented our approach on a clickstream dataset, which was collected from a Hungarian on-line news portal. Each clickstream transaction is of variable length. The presented clustering technique is useful in discovering the pattern of groups of users with similar interest and motivation for visiting a particular website. This study is also helpful in building-up adaptive web server depending on the users' behavior.

REFERENCES

Jhoshi, A. and Krishnapuram , R., " Robust fuzzy clustering methods to support web mining, proceedings of the workshop on Data Mining and Knowledge Discovery, SIGMOD ' 98, Seattle, pp. 15/1 – 15/8, June 1998.

Bezdek, J, C,. Pattern recognition and fuzzy objective function algorithms, plenum Press , New York 1981.

Hogo, M., Snorek, M. and Lingras, P., Temporal versus latest snapshot web usage mining using kohonen som and modified kohonen som based on the properties of rough sets theory , international journal on artificial intelligence tools, vol. 13, no. 3 (2004) 569-591.

Cooley, R., *Web Usage Mining: Discovery and Applications of Interesting Patterns from Web data*. PhD thesis,Dept. of Computer Science, University of Minnesota, May 2000.

Spiliopoulou, M. and Faulstich, L, C., WUM: A tool for web utilization analysis. In *Extended version of Proc.EDBT Workshop WebDB'98*, pages 184–203. Springer Verlag, 1999.

Mannila, H. and Meek, C., Global partial orders from sequential data. In *Proc. 6th Intl. Conf. on KnowledgeDiscovery and Data Mining (KDD2000)*, pages 161–168, Aug 2000.

Pitkow, J. and Pirolli, P., Mining longest repeating subsequences to predict world wide web surfing. In *Proc 2ndUSENIX Symposium on Internet Technologies & Systems (USITS'99)*, Oct 1999.

Fu, Y., Sandhu, K. and. Shih, M., A generalization-based approach to clustering of web usage sessions. In

Dash, M., Huan, L., Peter, S., KianLee, T.: Fast Hierarchical Clustering and its Validation, Data and Knowledge Engineering. 44(1) (2003) 109-138.

De, S, K., Radha Krishna, P.: Mining web data using clustering technique for web personalization, Int. Jour. of Computational Intelligence and Applications, 2(3) (2002) 255-265.

De, S.K., Radha Krishna, P.: Clustering web transactions using rough approximation, Fuzzy Sets and Systems (2004) (In print).

Jamil, S., Jitender, S,D. : Concept Approximations Based on Rough Sets and similarity Measures, International Journal on Applied Mathematics and Computer Science,2001,Vol.11,No.3,655 – 674.

Pawlak, Z., Rough Sets, International Journal of Computer and Information Sciences, 11 (1982) 341-356.

Perkowitz, M., Etzioni, O.: Towards adaptive web sites: Conceptual framework and case study, Artificial Intelligence, 118 (2000) 245-275.

Spiliopoulou, M.: Data mining for the web, In Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD'99, (1999) 588- 589.

Manco, G., Ortale, R., and Sacca, D., Similarity-based clustering of Web transactions, Symposium on Applied Computing**,** Proceedings of the 2003 ACM symposium on Applied computing pp. 1212 - 1216