ROUGH CLUSTERING FOR WEB TRANSACTIONS

IWAN TRI RIYADI YANTO

A thesis submitted in fullfillment of
the requirements for the award of the
Degree of Master of Information Technology

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

JANUARY 2011

ABSTRACT

Grouping web transactions into clusters is important in order to obtain better understanding of user's behavior. Currently, the rough approximation-based clustering technique has been used to group web transactions into clusters. It is based on the similarity of upper approximations of transactions by given threshold. However, the processing time is still an issue due to the high complexity for finding the similarity of upper approximations of a transaction which used to merge between two or more clusters. In this study, an alternative technique for grouping web transactions using rough set theory is proposed. It is based on the two similarity classes which is nonvoid intersection. The technique is implemented in MATLAB® version 7.6.0.324 (R2008a). The two UCI benchmark datasets taken from: http:/kdd.ics.uci.edu/ databases/msnbc/msnbc.html and http:/kdd.ics.uci.edu/databases/ Microsoft / microsoft.html are opted in the simulation processes. The simulation reveals that the proposed technique significantly requires lower response time up to 62.69 % and 66.82 % as compared to the rough approximation-based clustering, severally. Meanwhile, for cluster purity it performs better until 2.5 % and 14.47%, respectively.

ABSTRAK

Pengelompokan transaksi web kepada sesuatu kumpulan adalah penting untuk mendapatkan pemahaman yang lebih baik tentang perilaku pengguna. Pada masa kini, teknik pengelompokan berasaskan-anggaran kasar telah digunakan untuk mengelompokkan transaksi web kepada sesuatu kumpulan. Kaedah ini berdasarkan kepada kesamaan anggaran atas bagi sesuatu transaksi oleh satu nilai had tertentu. Walau bagaimanapun, masa pemprosesan masih menjadi satu isu disebabkan oleh kerumitan yang tinggi untuk mencari kesamaan atas bagi sesuatu transaksi yang digunakan untuk menggabungkan antara dua atau lebih kumpulan. Dalam kajian ini, suatu tenik alternatif untuk pengelompokan transaksi web menggunakan teori set kasar digunakan. Ianya adalah berdasarkan kepada dua kelas kesamaan yang mempunyai persilangan nonvoid. Teknik ini diimplementasi di dalam MATLAB® Versi 7.6.0.324 (R2008a). Dua data UCI yang diambil daripada: http:/kdd.ics.uci.edu/databases/msnbc/msnbc.html dan http:/kdd.ics.uci.edu/databases/Microsoft/microsoft.html digunakan dalam proses simulasi. Simulasi menunjukkan bahawa tenik yang dicadangkan secara signifikannya memerlukan masa respon yang lebih rendah, masing-masing sehingga 62.69% dan 66.82% jika dibandingkan dengan teknik pengelompokkan berasaskan-anggaran kasar. Manakala, prestasi kualiti kelompok adalah lebih baik di mana masing-masing mencapai sehingga 2.5% dan 14.47%.

PUBLICATIONS

A fair amount of the materials presented in this thesis has been published in various refereed conference proceedings and journals.

1. Iwan Tri Riyadi Yanto, Tutut Herawan and Mustafa Mat Deris. A framework on rough set approach for clustering web transactions. N.T. Nguyen et al. (Eds.): Advances in Intelligent Information and Database Systems, Studies in Computational Intelligence 283, pp. 265–277. © Springer-Verlag Berlin Heidelberg 2010.

2. Iwan Tri Riyadi Yanto, Tutut Herawan and Mustafa Mat Deris. RoCeT: rough set approach for clustering web transactions. Manuscript accepted in a special issue of Soft Computing Methodology, International Journal of Biomedical and Human Sciences, Japan, to appear Vol. 17, No. 1, July 2010.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background

Web data includes data from web server access logs, proxy server logs, browser logs, user profiles, registration files, user sessions or transactions, user queries, bookmark folders, mouse clicks and scrolls, and any other data generated by the interaction of users and the web (Pal, Talwar and Mitra, 2002). Web mining works on user profiles, user access patterns and mining navigation paths which are being heavily used by e-commerce companies for tracking customer behavior on their sites.

Web Usage Mining (WUM) is an active area of research and commercialization. Generally, web mining techniques can be defined as those methods to extract so called "nuggets" (or knowledge) from web data repository, such as content, linkage, usage information, by utilizing data mining tool. Among such web data, user click stream, i.e. usage data, can be mainly utilized to capture users' navigation patterns and identify user intended tasks. Once the user navigational behaviors are effectively characterized, they will provide benefit for further web applications, in turn, facilitate and improve web service quality for both web-based organizations and for end users (Bucher and Mulvena, 1998; Cohen, Krishnamurthy and Rexford, 1998; Joachims, Freitag and Mitchell, 1997; Lieberman, 1995; Mobasher, Cooley and Srivastava, 1999; Ngu and Wu, 1997; Perkowitz and Etzioni, 1998; Yanchun, Guandong, and Xiaofang, 2005). Recently, A number of approaches

have been developed for dealing with specific aspects of Web usage mining for the purpose of automatically discovering user profiles. 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. Schechter et al. (Schechter, Krishnan and Smith, 1998) have developed techniques for using path profiles of users to predict future HTTP requests, which can be used for network and proxy caching. Spiliopoulou et al. (Spiliopoulou, 1999) and Cooley (Cooley, 2000) have applied data mining techniques to extract usage patterns from web logs, for the purpose of deriving market intelligence. Hong et al. (Hong, Kuo and Chi, 1999) proposed a fuzzy mining approach for finding association rules from transactions.

The existing web usage data mining techniques include statistical analysis, association rules, sequential patterns, classification, and clustering. An important topic in Web Usage Mining is clustering web users – discovering clusters of users that exhibit similar information needs, e.g., users that access similar pages (Colley, Mobasher and Srivastava, 1999). It (Perkowitz and Etzioni, 2000; Han, 1998) is adopted widely to improve the usability and scalability of web mining. By analyzing the characteristics of the clusters, web users can be understood better and thus can be provided with more suitable and customized services (Colley, Mobasher and Srivastava, 1999). While calculating the similarity between two user transactions, considering a small amount of dissimilarity is meaningless. rough set theory can be applied to clustering web usage mining.

A well-known approach for clustering web transactions is using rough set theory (Pawlak, 1982; Pawlak 1991; Pawlak and Skowron, 2007). De and Krishna proposed an algorithm for clustering web transactions using rough approximation (De and Krishna, 2004). It is based on the similarity of upper approximations of transactions by given threshold. However, some iteration should be done to merge of two or more clusters that have the same similarity of upper approximations and didn't present how to handle the problem if, by given high threshold value, in the case a large number of user transactions, there are many clusters that are spread and the knowledge of remainder user behaviour cannot be known.

1.2 Problem statement

Recently, there are many previous researches that discuss about web usage mining. Discovery of user navigation pattern of Web page is important in order to identify potential customers for marketing strategy and build customized web services for individual users such as adaptive sites. Currently, less web usage mining model used clustering and most of the recent model used similarity as the parameter to merge between two clusters. Furthermore, they did not provide better solution in terms of lower processing time and less time complexity.

Clustering web transactions using rough approximation proposed by De and Khrisna is technique used to extract useful information from web transactions based on rough set theory. Similarity of upper approximation is used for clustering the clicks of user navigation (De and Krishna, 2004). However, the processing time is still an issue due the high computational complexity for finding the similarity of upper approximations of a transaction which used to merge between two or more clusters. Futhermore, The cluster purity is still an issue if there are so many clusters have only one transaction by given only one threshold.

1.3 Objectives and Scope

The objectives of this research are

a. To develop a technique for clustering web transaction using rough set theory based on non void intersection between two similar classes, achieves lower computation complexity and higher clusters purity.
b. To test the proposed technique on small datasets and benchmark datasets.

The scope of this research includes clustering access transaction over the web that can be expressed in two finite sets namely user behavior and hyperlink (URL). Two assumptions are made, firstly a user transaction set is a sequence of items that is formed by *m* users and secondly, an attribute set is made up of *n* distinct clicks (hyperlink s/URLs). Rough set theory is applied to determine the clustering technique.

1.4 Contributions

The contributions of this research are stated below.

a. Reducing complexity in clustering web transactions process.

b. Increasing clusters purity in classifying user web transactions.

1.5 Thesis Organization

The organization of the thesis is as follows. Chapter 2 describes reviews of the existing researches that are related to this research and the fundamental concept of rough set theory. Chapter 3 describes the fundamental concept of rough set theory. Chapter 4 describes the proposed techniques for clustering web transactions, referred to as similarity classes. We have proven that two similarity classes with non void intersection will be allocated in the same clusters. The experimental results of the proposed technique and its performance comparison with existing technique will be discussed and analyzed in Chapter 5. Finally, the conclusion and future work will be described in Chapter 6.

CHAPTER 2

LITERATUR REVIEW

This chapter describes and review existing researches that are related to web clustering using rough set theory research.

2.1 Clustering

Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields. In other words, data clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. Depending on the nature of the data and the purpose for which clustering is being used, different measures of similarity may be used to place items into classes, where the similarity measure controls how the clusters are formed. Some examples of measures that can be used as in clustering include distance, connectivity, and intensity.

An important step in any clustering is to select a distance measure, which will determine how the similarity of two elements is calculated. This will influence the

shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another.

2.2 Web

The World Wide Web is a system of interlinked hypertext documents accessed via the internet. With a web browser, one can view Web pages that may contain text, images, videos, and other multimedia and navigate between them using hyperlinks. The terms Internet and World Wide Web are often used in every day speech without much distinction. However, the Internet and the World Wide Web are not same. The Internet is a global system of interconnected computer networks. In contrast, the Web is one of the services that run on the Internet. It is a collection of interconnected documents and other resources, linked by hyperlinks and URLs. In short, the Web is an application running on the Internet.

Access transaction over the web can be expressed in the two finite sets, user transaction and hyperlinks/URLs. A user transaction set *U* is a sequence of items. This set is formed by m users. Set *A* is set of distinct *n* clicks (hyperlinks/URLs), clicked by users that are

$$U = \{t_1, t_2, \ldots, t_m\},$$
$$A = \{hl_1, hl_2, \ldots, hl_n\}.$$

Where, $\forall t_i \in U$ is a non-empty subset of *U*. The temporal order of user's clicks within transactions has been taken into account. A user transaction $t \in U$ is represented as a vector:

$$t = [u_1^t, u_2^t, \ldots, u_n^t],$$

Where

$$u_i^t = \begin{cases} 1 & \text{if} \quad \text{hl}_\text{i} \in t, \\ 0 & \text{otherwise.} \end{cases}$$

Example 2.1. In this example, the data set in (De and Krishna, 2004) is used. De and Krishna provide a web user transaction containing four objects $(|U|=4)$ with five hyperlinks$(|A|=5)$. Let $U=\{t_1,t_2,t_3,t_4\}$ be the set of User transactions and $A=\{hl_1,hl_2,hl_3,hl_4,hl_5\}$ be a set of distinct URLs accessed from the user transaction *U*. Let

$$t_1=\{hl_1,hl_2\},$$
$$t_2=\{hl_2,hl_3,hl_4\},$$
$$t_3=\{hl_1,hl_3,hl_5\},$$
$$t_4=\{hl_2,hl_3,hl_5\}.$$

These user transactions can be represented as vectors

$$t_1=\{1,1,0,0,0\},\quad t_2=\{0,1,1,1,0\},\quad t_3=\{1,0,1,0,1\},\quad t_4=\{0,1,1,0,1\}.$$

2.3 Web usage mining

Web usage mining is the application of data mining techniques to web log data repositories. Discovering user access patterns from web access log is increasing the importance of information to build up adaptive web server according to the individual user's behavior. Web usage mining tries to make sense of the data gathered by the web surfer's behaviors (Cooley, Mobasher and Srivastava, 2000). While the web content and structure mining utilize the real or primary data on the web, web usage mining uses the secondary data derived from the interactions of the users while interacting with the web. Web usage data includes the data from web server access logs, proxy server logs, user profiles, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks, scrolls and any other data as results of the interaction.

Existing web usage data mining techniques include statistical analysis (Srivastava, Cooley and Deshpande, 2000), association rules (Mobasher, Dai and Nakawaga, 2001; Huang, Cercone and Promhouse, 2002), sequential patterns (Yang and Zhang, 2001), classification (Yang and Wang, 2001), and clustering (Mobasher, Cooley and Srivastava, 1999; Labroche, Monmarche and Venturini, 2003). An important topic in Web Usage Mining is clustering web users – discovering clusters of users that exhibit similar information needs, e.g., users that access similar pages (Ling, Brownick and Nejdl, 2009). It is useful to find interesting user access patterns in web log (De and Krishna, 2004).

2.4 Web user clustering

.

Web clustering is the unsupervised classification of patterns into groups. The clustering problem is partitioning a user transaction into clusters. The user transaction is a set of *n* elements described by *m* attributes. The goal is to group a set of patterns into a number of more or less homogeneous clusters with respect to a suitable similarity measure. Patterns that are similar are allocated in the same cluster, while the patterns that differ significantly are put in different clusters.

There are quite a few methods for clustering web users proposed in the literature (Fu, Sandhu and Shih, 1999; Xiao and Zhang, 2001; Wang, 2002). In general, web user clustering consists of three phases: data preparation, cluster discovery, and cluster analysis. Since the last phase is application-dependent, let us briefly describe the first two. In the first phase, web sessions of users are extracted from the web server log by using some user identification and session identification techniques (Cooley, Mobasher, Srivastava, 1999). A web session, which is an episode of interaction between a web user and the web server, consists of pages visited by a user in the episode (Fu, Sandhu and Shih, 1999). In the second phase, clustering techniques are applied to generate clusters of users.

Many clustering algorithms have been presented in Refs. (Jain, Murty and Flynn, 1999; Xu and Wunsh, 2005), including K-Means (McQueen, 1967), agglomerative hierarchical clustering (Day and Edelsbrunner, 1985), graph partitioning (Zahn, 1971) and spectral clustering (Hea, Zhaa, Ding and Simon, 2002), density-based methods (Ester, Kriegel, Sander and Xu, 1996), grid-based methods,

and model-based methods. So far, agglomerative hierarchical clustering (AHC), K-Means, and some graph partitioning methods have been employed for most web clustering applications. Among these algorithms, AHC is a well-known hierarchical method that combines the smaller data items to create larger clusters. An AHC algorithm starts with each web page in a single cluster and merges the two most similar clusters iteratively until a halting criterion is satisfied. The measure used for combination can be the nearest, the average, and the farthest distances, and accordingly these methods can be classified into several categories (Day and Edelsbrunner, 1985): single linkage, average linkage, and complete linkage. The halting criterion in AHC is based on a predetermined constant (Miligan and Cooper, 1985). Because of sensitivity of AHC to the halting criterion, the algorithm may mistakenly merge two good clusters. Although hierarchical methods are often said to have better quality clustering results (Jain and Murty, 1999), they usually do not reallocate the pages that may have been poorly clustered in the early stages of the process (Zhao and Karypis, 2005). Moreover, the time complexity of hierarchical methods is quadratic (Stenbach, 2000).

The web user clustering results provide knowledge on users' information needs and can be used in various web personalization applications. It discovers knowledge on the evolutionary characteristics of users' information needs.

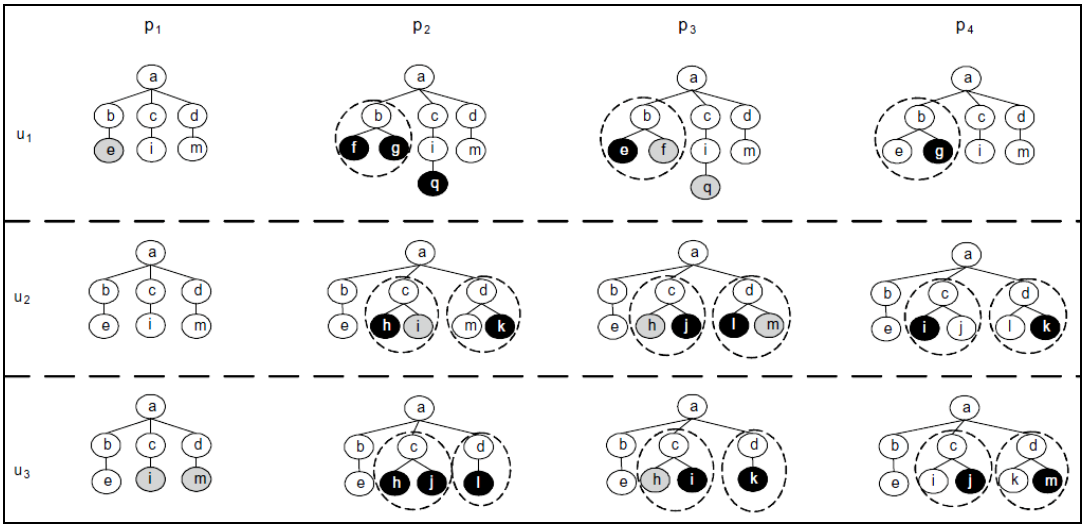

Figure 2.1: The historical web users.

For example, given the web users and their historical web sessions in Figure 2.1, it is easy to know that u_2 and u_3 will be clustered together while leaving u_1 in

another cluster. Knowledge can be inferred from the clustering result that users in the same cluster exhibit similar variation patterns in their information needs. Thus, the results of web user clustering are useful for various web personalization scenarios. The some applications of web user clustering in the following paragraphs (Chen, Bhowmick and Nejdl, 2009).

a. Intelligent web advertisement

Almost all of web sites offer standard banner advertisements (Buchwalter, Ryan and Martin, 2001), underlying the importance of this form of on-line advertising. For many web-based organizations, revenue from advertisements is often the only or the major source of income (example, Yahoo.com, Google.com) (Amiri and Menon, 2003). One of the ways to maximize revenues for the party who owns the advertising space is to design intelligent techniques for the selection of an appropriate set of advertisements to display in appropriate web pages. Web user clustering can be beneficial for designing intelligent advertisement placement strategies.

b. Proxy cache management

Web caching is another interesting problem (Cao and Irani, 1997; Yang and Zang, 2001) as web caches can reduce not only network traffic but also downloads latency. Because of the limited size of cache regions, it is important to design effective replacement strategies to maximize hit rates. One of the frequently used replacement strategies is LRU (Least Recently Used), which assigns priorities to the most recently accessed pages. Web user clusters can be used with LRU to manage the caching region more optimally (Chen, Bhowmick and Nejdl, 2009).

2.5 Clustering web transaction using rough approximation

A web site is usually rich in information content and is complicated in hyperlink structure, which allows users to navigate. Recently, a number of approaches have been proposed in clustering web usage mining. Adaptive data mining techniques have been successfully applied on web log data (Chen, Park and Yu, 1998; Ng and Chan, 2001; Pei, Han and Mortazaviasl, 2000; Perkowitz and Etzioni, 2000). The usage patterns of the web are different for different users and also the same user navigates the same pattern in different ways. In general, the discovered knowledge or

any unexpected rules are likely to be imprecise or incomplete. While calculating the similarity between two user transactions, considering a small amount of dissimilarity is meaningless. rough set theory can be applied to clustering web usage mining. There has been work in the area of applying rough set theory to the process of clustering web user transactions including (De and Khrisna, 2004).

De and Krishna proposed a method of clustering the clicks of user navigations called as Similarity Upper Approximation. The technique (De and Khrisna, 2004) needs three main steps. The first step of the technique is obtaining the measure of similarity that gives information about the users access patterns related to their common areas of interest by similarity relation between two transactions of objects.

Definition 2.1. Given two transactions *t* and *s*, the measure of similarity between *t* and *s* is given by:

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

Obviously $sim(t,s) \in [0,1]$, where $sim(t,s) = 1,$ when two transactions *t* and *s* are exactly identical, and $sim(t,s) = 0,$ when two transactions t and s have no items in common.

Second step is to obtain the similarity classes. A binary relation *R* defined on U is used. Given user-defined threshold value $\text{th} \in [0,1]$ and for any two user transactions *t* and $\text{s} \in U$, a binary relation *R* on *U* denoted as *tRs* is defined by *tRs* iff $sim(t,s) \geq \text{th}$. This relation R is a tolerance relation as *R* is both reflexive and symmetric but transitive may not hold good always.

Definition 2.2. The similarity class of *t*, denoted by *R*(*t*), is the set of transactions which are similar to *t*. It is given by

$$R(t) = \{\text{s} \in |sRt\}.$$

For different threshold values we can get different similarity classes. A domain expert can choose the threshold based on his experience to get a proper similarity

class. It is clear that for a fixed threshold $\in [0,1]$ a transaction form a given similarity class may be similar to an object of another similarity class. Since the similarity classes will be obtained iff $sim(t,s) \geq$ th, thus the threshold should be less then the highest the measure of similarity.

The third step is to cluster the transactions based on similarity upper approximations(De and Khrisna, 2004).

Definition 2.3. Let $P \subset U$, for a fixed threshold $\in [0,1]$, a binary tolerance relation R is defined on U. The lower approximation of P, denoted by $\underline{R}(P)$ and the upper approximation of P, denoted by $\overline{R}(P)$ are respectively defined as follows:

$$\underline{R}(P) = \{t \in P : R(t) \subseteq P\} \text{ and } \overline{R}(P) = \bigcup_{t \in P} R(t),$$

Obviously, let's $t_i \in U$ be a user click-stream. $\overline{R}(t_i)$ is a set of transactions similar to t_i i.e. 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 similar to $\overline{R}(t_i)$. This process continues until two consecutive upper approximations for $t_i, i = 1,2,3,\ldots,|U|$ are same and two or more clusters that have the same similarity upper approximations merges at each iteration. This can be called the Similarity Upper Approximation of transaction $t_i \in U$, denoted S_i .

2.6 Summary

In summary, from the discussion of this chapter, several models of clustering algorithm on web usage mining have been introduced. These models are designed to enhance the performance from the previous web usage mining model. However, one needs to take into consideration the processing time as well as time complexity and purity of the clusters.

CHAPTER 3

PRELIMINARIES

3.1 Rough Set Theory

The observation that one cannot distinguish objects on the basis of given information about them is the starting point of rough set theory. In other words, imperfect information causes indiscernibility of objects. The indiscernibility relation induces an approximation space made of equivalence classes of indiscernible objects. A rough approximating a subset of the set of objects is a pair of dual set approximations, called a lower and an upper approximation in term of these equivalence classes (Pawlak, 1982; Pawlak 1991; Pawlak and Skowron, 2007). Rough sets are defined through their dual set approximations in an information system. The notion of information system provides a convenient tool for the representation of objects in terms of their attribute values.

3.1.1 Information system

An information system is a 4-tuple (quadruple) $S = (U, A, V, f)$, where $U = \{u_1, u_2, u_3, \ldots, u_{|U|}\}$ is non-empty finite set of objects, $A = \{a_1, a_2, a_3, \ldots, a_{|A|}\}$ is

non-empty finite attributes, $V = \bigcup_{a \in A} V_a$, V_a is the domain (value set) of attribute *a*, $f : U \times A \to V$ is an function such that $f(u, a) \in V_a$, for every $(u, a) \in U \times A$, called information (knowledge) function.

Table 3.1: An information system

U	a_1	a_2	…	a_k	…	$a_{\|A\|}$
u_1	$f(u_1, a_1)$	$f(u_1, a_2)$	…	$f(u_1, a_k)$	…	$f(u_1, a_{\|A\|})$
u_2	$f(u_2, a_1)$	$f(u_2, a_2)$	…	$f(u_2, a_k)$	…	$f(u_2, a_{\|A\|})$
⋮	⋮	⋮	⋱	⋮	⋱	⋮
$u_{\|U\|}$	$f(u_{\|U\|}, a_1)$	$f(u_{\|U\|}, a_2)$	…	$f(u_{\|U\|}, a_k)$	…	$f(u_{\|U\|}, a_{\|A\|})$

In many applications, there is an outcome of classification that is known. This *a posteriori* knowledge is expressed by one (or more) distinguished attribute called decision attribute. The process is known as supervised learning. An information system of this kind is called a decision system. A *decision system* is an information system of the form $D = (U, A = C \cup D, V, f)$, where *D* is the set of *decision attributes* and $C \cap D = \phi$. The elements of *C* are called *condition attributes*. A simple example of decision system is given in Table 3.2.

Example 3.1. Suppose that the data about 6 people is given in Table 3.2.

Table 3.2: A student decision system

Student	**Diploma**	**Experience**	**Reference**	**Decision**
1	Degree	medium	Neutral	reject
2	Phd	high	Good	accept
3	MSc	medium	Neutral	accept
4	MSc	low	Excellent	accept
5	Degree	high	Good	accept
6	MSc	medium	Neutral	reject

The following values are obtained from Table 3.2,

$U = \{1,2,3,4,5,6\}$,

$A = \{\text{Diploma, Experience, Reference, Decision }\}$, where

$C = \{\text{Diploma, Experience, Reference }\}$, $D = \{\text{Decision}\}$

$V_{\text{Diploma}} = \{\text{Degree, MSc, Phd}\}$,

$V_{\text{Experience}} = \{\text{low, medium, high}\}$,

$V_{\text{Reference}} = \{\text{Neutral, Good, Excellent}\}$,

$V_{\text{Decision}} = \{\text{accept, reject}\}$.

A relational database may be considered as an information system in which rows are labeled by the objects (entities), columns are labeled by attributes and the entry in row *u* and column *a* has the value $f(u,a)$. It is noted that each map $f(u,a): U \times A \to V$ is a tupple $t_i = \left(f(u_i,a_1), f(u_i,a_2), f(u_i,a_3), \cdots, f(u_i,a_{|A|})\right)$, for $1 \le i \le |U|$, where $|X|$ is the cardinality of *X*. Note that the tuple *t* is not necessarily associated with entity uniquely (refers to students 3 and 6 in Table 3.2). In an information table, two distinct entities could have the same tupple representation (duplicated/redundant tupple), which is *permissible* in relational databases. Thus, the concepts in information systems are a generalization of the same concepts in relational databases.

3.1.2 Set Approximations

The starting point of rough set approximations is the indiscernibility relation, which is generated by information about object of interest. Two objects in an information system are called indiscernible (indistinguishable or similar) if they have the same feature.

Definition 3.1. *Two elements $x, y \in U$ are said to B-indiscernible (indiscernible by the set of attribute $B \subseteq A$ in S) if and only if $f(x,a) = f(y,a)$, for every $a \in B$.*

Obviously, every subset of *A* induces unique indiscernibility relation. Notice that, an indiscernibility relation induced by the set of attribute *B*, denoted by *IND*(*B*), is an equivalence relation. The partition of *U* induced by *IND*(*B*) is denoted by *U/B* and the equivalence class in the partition *U/B* containing $x \in U$, is denoted by $[x]_B$. The notions of lower and upper approximations of a set are defined as follow.

Definition 3.2. The B-lower approximations of X, denoted by $\underline{B}(X)$ and B-upper approximation of X, denoted by $\overline{B}(X)$, respectively, are defined as.

$$\underline{B}(X) = \{x \in U \mid [x]_B \subseteq X\} \text{ and } \overline{B}(X) = \{x \in U \mid [x]_B \cap X \neq \phi\}.$$

From Definition 3.2, the following interpretations are obtained

a. The *lower approximation* of a set *X* with respect to *B* is the set of all objects, which can be for *certain* classified as *X* using *B* (are certainly *X* in view of *B*).
b. The *upper approximation* of a set *X* with respect to *B* is the set of all objects which can be *possibly* classified as *X* using *B* (are possibly *X* in view of *B*).

Hence, with respect to arbitrary subset $X \subseteq U$, the universe *U* can be divided into three disjoint regions using the lower and upper approximations

a. The *positive region* $\mathrm{POS}_B(X) = \underline{B}(X)$, i.e., the set of all objects, which can be for *certain* classified as *X* using *B* (are *certainly X* with respect to *B*).
b. The *boundary region* $\mathrm{BND}_B(X) = \overline{B}(X) - \underline{B}(X)$, i.e., the set of all objects, which can be classified neither as *X* nor as not-*X* using *B*.
c. The *negative region* $\mathrm{NEG}_B(X) = U - \overline{B}(X)$, i.e., the set of all objects, which can be for *certain* classified as not-*X* using *B* (are *certainly* not-*X* with respect to *B*).

Example 3.2. Let us depict above notions by examples referring to Table 3.2.
Let attribute subset $B = \{Diploma\}$, given the arbitrary set $X = \{1,2,5\}$. Since the equivalence classes of $U / B = \{\{1,5\}, \{2\}, \{3,4,6\}\}$, the lower and upper approximation the attribute B respect to the target set X are

$$\underline{B}(X) = \{2\} \text{ and } \overline{B}(X) = \{1,5\} \cup \{2\} = \{1,2,5\}.$$

Thus, for this case, three disjoint regions the attribute B respect to the target set $X = \{1,2,5\}$, is given as

$$\mathrm{POS}_B(X) = \underline{B}(X) = \{2\},$$

$$\mathrm{BND}_B(X) = \overline{B}(X) - \underline{B}(X) = \{1,5\},$$

$$\mathrm{NEG}_B(X) = U - \overline{B}(X) = \{3,4,6\}.$$

These notions of lower and upper approximations can be shown clearly as in Figure 3.1.

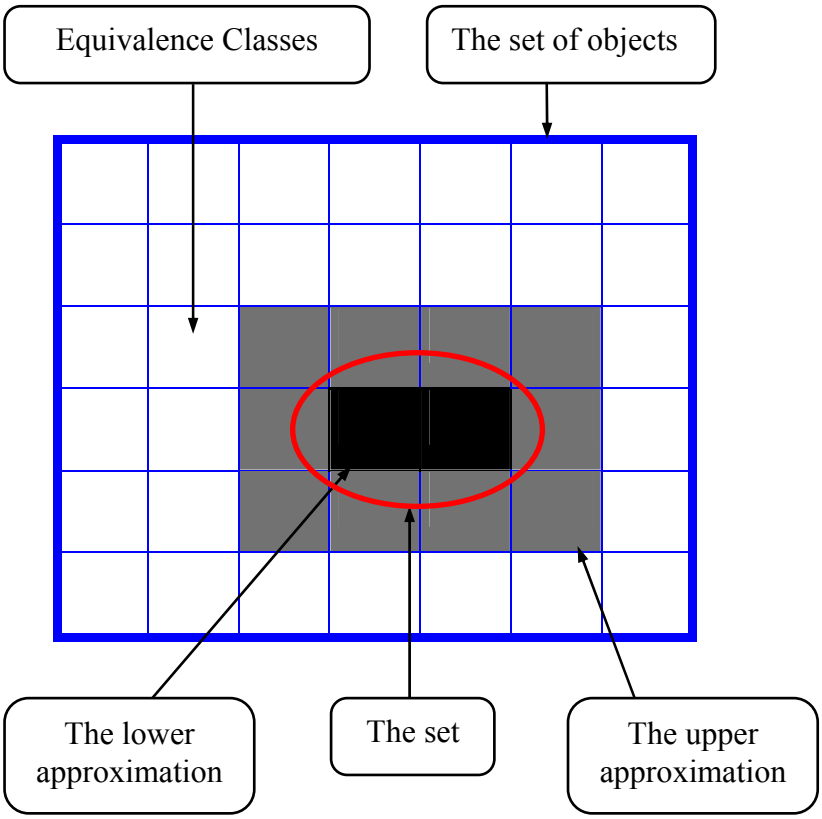

Figure 3.1: Set approximations

Let ϕ be the empty set, $X, Y \subseteq U$ and $\neg X$ be the complement of X in U. The lower and upper approximations satisfy the following properties (Zhu, 2007):

(1L)	$\underline{B}(U) = U$	(Co-Normality)
(1U)	$\overline{B}(U) = U$	(Co-Normality)
(2L)	$\underline{B}(\phi) = \phi$	(Normality)
(2U)	$\overline{B}(\phi) = \phi$	(Normality)

(3L)	$\underline{B}(X) \subseteq X$	(Contraction)
(3U)	$X \subseteq \overline{B}(X)$	(Extension)
(4L)	$\underline{B}(X \cap Y) = \underline{B}(X) \cap \underline{B}(Y)$	(Multiplication)
(4U)	$\overline{B}(X \cup Y) = \overline{B}(X) \cup \overline{B}(Y)$	(Addition)
(5L)	$\underline{B}(X \cup Y) \supseteq \underline{B}(X) \cup \underline{B}(Y)$	(Inclusion)
(5U)	$\overline{B}(X \cap Y) \subseteq \overline{B}(X) \cap \overline{B}(Y)$	(Inclusion)
(6L)	$\underline{B}(\underline{B}(X)) = \underline{B}(X)$	(Idempotency)
(6U)	$\overline{B}(\overline{B}(X)) = \overline{B}(X)$	(Idempotency)
(7L)	$\underline{B}(\neg X) = \neg\overline{B}(X)$	(Duality)
(7U)	$\overline{B}(\neg X) = \neg\underline{B}(X)$	(Duality)
(8L)	$X \subseteq Y \Rightarrow \underline{B}(X) \subseteq \underline{B}(Y)$	(Monotone)
(8U)	$X \subseteq Y \Rightarrow \overline{B}(X) \subseteq \overline{B}(Y)$	(Monotone)
(9L)	$\underline{B}(\neg\underline{B}(X)) = \neg\underline{B}(X)$	(Lower Complement Relation)
(9U)	$\overline{B}(\neg\overline{B}(X)) = \neg\overline{B}(X)$	(Upper Complement Relation)
(10L)	$\forall G \in U/B, \underline{B}(G) = G$	(Granurality)
(10U)	$\forall G \in U/B, \overline{B}(G) = G$	(Granurality)

The accuracy of approximation (accuracy of roughness) of any subset $X \subseteq U$ with respect to $B \subseteq A$, denoted $\alpha_B(X)$ is measured by

$$\alpha_B(X) = \frac{|\underline{B}(X)|}{|\overline{B}(X)|},$$

where |X| denotes the cardinality of X. For empty set ϕ, we define $\alpha_B(\phi) = 1$. Obviously, $0 \le \alpha_B(X) \le 1$. If X is a union of some equivalence classes of *U*, then $\alpha_B(X) = 1$. Thus, the set X is crisp (precise) with respect to *B*. and, if X is not a union of some equivalence classes of *U*, then $\alpha_B(X) < 1$. Thus, the set X is rough (imprecise) with respect to *B*. This means that the higher of accuracy of approximations of any subset $X \subseteq U$ is more precise (the less imprecise) of itself.

Example 3.3. Table 3.2 shows a student decision system dataset. Consider the concept "Decision", i.e., the set $X(\text{Decision} = \text{accept}) = \{2,3,4,5\}$ and the set of attributes $C = \{\text{Diploma}, \text{Experience}, \text{Reference}\}$. The partition of U induced by $IND(C)$ is given by

$$U / C = \{\{1\},\{2\},\{3,6\},\{4\},\{5\}\}.$$

The corresponding lower approximation and upper approximation of X are as follows

$$\underline{C}(X) = \{2,4,6\} \text{ and } \overline{C}(X) = \{2,3,4,5,6\}.$$

Thus, concept "Decision" is imprecise (rough). For this case, the accuracy of approximation is given as

$$\alpha_C(X) = \frac{3}{5}.$$

It means that the concept "Decision" can be characterized partially employing attributes Diploma, Experience, Reference.

3.2 Summary

In this chapter, the concept of rough set theory through data contained in an information system has been presented. The rough set approach seems to be of fundamental importance to artificial intelligent, especially in the areas of decision analysis and knowledge discovery from databases (Pawlak, 1997; Pawlak, 2002; Pawlak, 2002). Basic ideas of rough set theory and its extensions, as well as many interesting applications can be found in (Peters and Skowron, 2006; http://roughsets.home.pl/www/; http://en.wikipedia.org/wiki/Rough_set).

CHAPTER 4

ROUGH CLUSTERING FOR WEB TRANSACTIONS

In this chapter, the technique for clustering web transaction using rough set theory which refer to as rough clustering for web transactions (ROCET) is proposed.

4.1 Grouping into the cluster

The technique for clustering the transactions is based on all the possible similarity of the similarity classes. The union of two similarity classes with non void intersection will be in the same clusters. The justification that a cluster is a union of two similarity classes with non-void intersection is presented in Proposition 4.2.

Definition 4.1. *Two clusters* Cl_i and $Cl_j, i \neq j$ *are said to be the same cluster, if*

$$\begin{aligned} Cl_j &= R(t_j) \bigcup R(t_i) \\ &= Cl_i \\ &= \bigcup R(t_i), i = 1,2,3,\cdots,|U| \end{aligned}$$

Proposition 4.2. *Let* Cl_i *and* Cl_j *be two clusters. If* $\bigcap R(t_i) \neq \phi$ *, then* $\bigcup R(t_i) = Cl_i$.

Proof. Suppose $Cl_i \neq Cl_j$, for $i \neq j$ we have $Cl_i \bigcap Cl_j = \phi$,

From Definition 4.1, If $\bigcup R(t_i) \neq Cl_i$, then $\bigcup R(t_i) = Cl_j$ and $\bigcup R(t_j) = Cl_i$, then

$$\left(\bigcup R(t_j)\right) \bigcap \left(\bigcup R(t_i)\right) = \phi,$$

$$\bigcap R(t_i) = \phi.$$

This is a contradiction from the hypothesis. □

4.2 Design of cluster selection

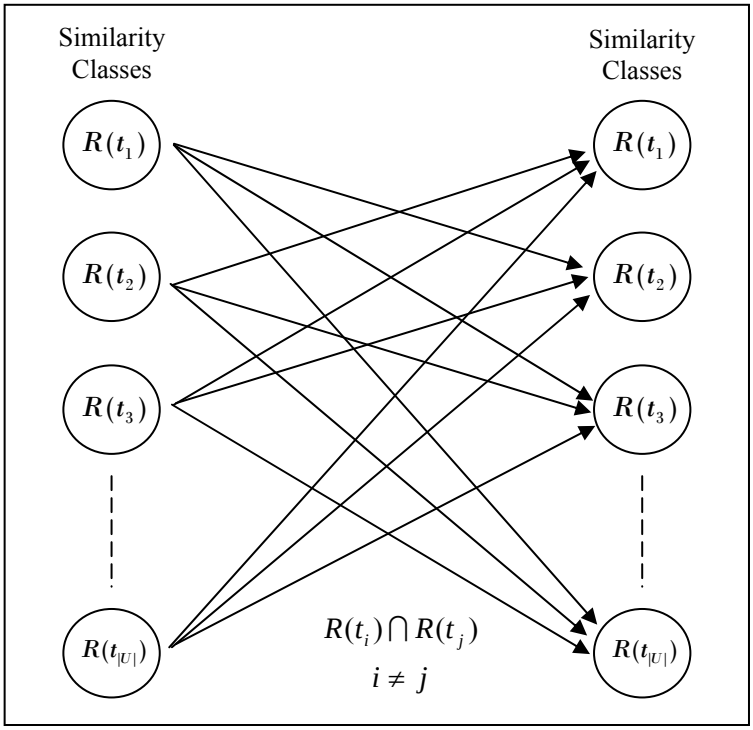

Figure 4.1: Design of cluster selection

The clustering decision is based on the two similarity classes with no empty intersection.

$$Cl_i = \{\bigcup R(t_i) \mid \bigcap R(t_i) \neq \phi\}.$$

Definition 4.3. Let $T \subset U$, for a fixed threshold $\in [0,1]$, a binary tolerance relation R is defined on U. Base on Definition 3.2, the lower approximation of T, denoted by $\underline{R}(T)$ and the upper approximation of T, denoted by $\overline{R}(T)$ are respectively defined as follows:

$$\underline{R}(T) = \{t \in U : R(t) \subseteq T\} \text{ and } \overline{R}(T) = \{t \in U : \bigcap R(t) \neq \phi\}.$$

Obviously, the Upper approximation is as the cluster.

Figure 4.2 shows the pseudo-code of ROCET algorithm. The algorithm uses the non void intersection between two similarity classes in information systems of web transactions. The algorithm consists of three main steps. The first step of the technique is obtaining the measure of similarity that gives information about the user access patterns related to their common areas of interest by similarity relation between two transactions of objects. The second step is obtaining of the Similarity classes by given the threshold value using Definition 2.2. The last step is to cluster the transaction, if the intersection between two similarity classes is non-empty, then they will be in the same cluster.

```
Algorithm: ROCET
Input: web transaction data set.
Output: Cluster of web transactions
Begin
        Step 1. Compute the measure of similarity between two
        transactions of objects.
        Step 2. Obtain the similarity classes by given threshold value.
        Step 3. Cluster the transactions if two of similarity classes have
        non-void   intersection.
End
```

Figure 4.2: ROCET algorithm

4.3 The computational complexity

Suppose that there are *n* objects of a web user transaction in an information system $S(U, A, V, f)$ of user transaction. Therefore, there are at most *n* similarity classes. Based on the De and Khrisna and ROCET techniques, n^2 computation are needed to determine the similarity matrix. However, the technique of De and Khrisna, the computation of Similarity Upper Approximations having *n* similarity classes t_i with respect to t_j, where $i \neq j$ is $n\,(n-1)$ times for each iteration until two consecutives upper approximations for t_i are the same. Assume that, the algorithm (De and Khrisna, 2004) converges after *m* iterations, where $m > 1$. Thus, the computational complexity of the De and Khrisna technique is $O(n^2 + m(n(n-1))),\quad m > 1$. Meanwhile, ROCET technique needs $n\,(n-1)$times to determine the union of two similarity classes. Thus, the computational complexity for ROCET technique is $O(n^2 + n(n-1))$.

Example 4.1. Please refer to web transaction in example 2.1 to obtain the measurement of similarity and the similarity classes.
The vector of user transactions in example 2.1 can be transformed in information system table as follows

Table 4.1: Data transactions

U/A	hl_1	hl_2	hl_3	hl_4	hl_5
t_1	1	1	0	0	0
t_2	0	1	1	1	0
t_3	1	0	1	0	1
t_4	0	1	1	0	1

The measures of similarity for web transactions from Table 4.1 are given bellow.

$$Sim(t_1, t_2) = \frac{|\, t_1 \cap t_2 \,|}{|\, t_1 \cup t_2 \,|} = \frac{|\, \{hl_2\} \,|}{|\, \{hl_1, hl_2, hl_3, hl_4\} \,|} = 0.25,$$

$$Sim(t_1,t_3)=\frac{|t_1\cap t_3|}{|t_1\cup t_3|}=\frac{|\{hl_1\}|}{|\{hl_1,hl_2,hl_3,hl_5\}|}=0.25,$$

$$Sim(t_1,t_4)=\frac{|t_1\cap t_4|}{|t_1\cup t_4|}=\frac{|\{hl_2\}|}{|\{hl_1,hl_2,hl_3,hl_5\}|}=0.25,$$

$$Sim(t_2,t_3)=\frac{|t_2\cap t_3|}{|t_2\cup t_3|}=\frac{|\{hl_3\}|}{|\{hl_1,hl_2,hl_3,hl_4,hl_5\}|}=0.2,$$

$$Sim(t_2,t_4)=\frac{|t_2\cap t_4|}{|t_2\cup t_4|}=\frac{|\{hl_2,hl_3\}|}{|\{hl_2,hl_3,hl_4,hl_5\}|}=0.5,$$

$$Sim(t_3,t_4)=\frac{|t_3\cap t_4|}{|t_3\cup t_4|}=\frac{|\{hl_3,hl_5\}|}{|\{hl_1,hl_2,hl_3,hl_5\}|}=0.5.$$

By setting the threshold to 0.5, then

$$Sim(t_1,t_2)<0.5,\qquad Sim(t_1,t_3)<0.5,\qquad Sim(t_1,t_4)<0.5,$$
$$Sim(t_2,t_3)<0.5,\qquad Sim(t_2,t_4)\geq 0.5,\qquad Sim(t_3,t_4)\geq 0.5.$$

Therefore, we will get the similarity classes as shown in figure 4.3.

$$R(t_1)=\{t_1\},\quad R(t_2)=\{t_2,\ t_4\},\quad R(t_3)=\{t_3,\ t_4\},\ R(t_4)=\{t_2,\ t_3,\ t_4\}.$$

Figure 4.3: The similarity classes by threshold 0.5

Hence, different similarity classes are obtained if the similarities web transactions are given different threshold.

To get the cluster (De and Khrisna, 2004), Similarity Upper Approximations is used. In the upper approximation processes for $\overline{R}(t_2)$ and $\overline{R}\overline{R}(t_2)$ are shown as follows.

a. To obtain $\overline{R}(P)$, since $\overline{R}(t_2)$, then $P=\{t_2\}$.

$$R(t_1)\Rightarrow t_1\notin P,\qquad R(t_2)\Rightarrow t_2\in P,$$
$$R(t_3)\Rightarrow t_3\notin P,\qquad R(t_4)\Rightarrow t_4\notin P,$$
$$\overline{R}(t_2)=\bigcup R(t_2)=\{t_2,t_4\}$$

REFERENCES

Amiri, S. Menon. (2003). *Efficient scheduling of internet banner advertisements*. ACM TOIT 3 (4) 334–346.

Bucher, A.G. and Mulvenna, M.D. (1998). *Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining*. SIGMOD Record, 27 (4), 54-61.

Buchwalter, C. Ryan, M. Martin, D. (2001). *The state of online advertising: data covering 4th Q 2000*. in TR Adrelevance.

Cao, P. Irani, S. (1997). *Cost-aware WWW proxy caching algorithms*. in: Proc. of USENIX SITSY.

Chehreghani, M. H. Abolhassani, H. Chehreghani, M. H. (2008). *Improving density-based methods for hierarchical clustering of web pages*. Data & Knowledge Engineering 67 30–50.

Chen, L. Bhowmick, S.S. Nejdl, W. (2009). *COWES: Web user clustering based on evolutionary web sessions*. Data & Knowledge Engineering 68 867–885.

Chen, J.S. Park, P.S. Yu. (1998). *Efficient data mining for traversal patterns*. IEEE Trans. Knowledge Data Eng. 10 (2) 209–221.

Cohen, E., Krishnamurthy, B. and Rexford, J. (1998). *Improving and-to-end performance of the web using server volumes and proxy lters*. Proceeding of the ACM SIGCOMM '98. Vancouver, British Columbia, Canada: ACM Press.

Cooley, R. (2000). *Web usage mining: discovery and application of interesting patterns from web data*. Ph. D. Thesis, Department of Computer Science, University of Minnesota, May 2000.

Cooley, R. Mobasher, B. Srivastava, J. (1999). *Data preparation for mining World Wide Web browsing patterns.* Journal of Knowledge and Information Systems 1 (1), 5 – 32.

Day, W.H.E. Edelsbrunner, H. (1985). *Investigation ofproportional link linkage clustering methods*. Journal of Classification, vol. 2, Springer-Verlag, New York Inc., pp. 239–254.

De, S.K. and. Krishna, P.R (2004). *Clustering web transactions using rough approximation*. Fuzzy Sets and Systems, 148, 131-138.

Ester, M. Kriegel, H.-P. Sander, J. Xu. X. (1996*). Adensity-based algorithm fordiscovering clusters inlarge spatial databases with noise*. KDD'96 226–231.

Fu, Y. Sandhu, K. Shih, M. (1999). *A generalization-based approach to clustering of web usage sessions.* in: Proc. of WEBKDD'99.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). *On clustering validation techniques.* Journal of Intelligent Information Systems 17 (2–3), 107–145.

Han, E. et al., (1998). *Hypergraph Based Clustering in High-Dimensional Data Sets: A Summary of Results*. IEEE Data Engineering Bulletin, 21 (11), 15-22.

Hea, X. Zhaa, H. Ding, C.H.Q. Simon, H.D. (2002). *Web document clustering using hyperlink structures*. Computational Statistics & Data Analysis 41 (1) 19–45.

Herawan, T. and Deris, M.M. (2009). *A direct proof of every rough set is a soft set*. In the Proceeding of International Conference of AMS 2009, IEEE Press, 119−124.

Hong, T.P. Kuo, C.S. Chi, S.C. (1999). *A data mining algorithm for transaction data with quantitative values.* Intelligent Data Anal. 3 (5) 363–376.

http:/kdd.ics.uci.edu/databases/ Microsoft / microsoft.html

http:/kdd.ics.uci.edu/ databases/msnbc/msnbc.html

Huang, X. An, A. Cercone, N. Promhouse, G. (2002). *Discovery of interesting association rules from livelink web log data*. in: Proc. of ICDM.

Joachims, T., Freitag, D. and Mitchell, T. Webwatcher, (1997). *A tour guide for the world wide web*. In the 15th international Joint Conference on Artificial Intelligence (ICJAI'97), Nagoya, Japan.

Labroche, N. Monmarche, N. Venturini, G. (2003*). Web sessions clustering with artificial ants colonies*. in: Proc. of www.2003.

Lieberman, H. Letizea. (1995). *An agent that assists web browsing*. Proceeding of the 1995 International Joint Conference on Artificial Intelligence. Montreal, Canada: Morgan Kaufmann.

Li, T. Yang, Q. Wang, K. (2001). *Classification pruning for web-request prediction.* in: Proc. of www, 2001.

McQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. in: Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297.

Milligan, G.W. Cooper, M.C. (1985). *An examination of procedures for detecting the number of clusters in a data set.* Psychometrika 50 159–179.

Mobasher, B. Cooley, R. Srivastava, J. (1999). *Creating adaptive web sites through usage-based clustering of URLs*, in: Proc. of IEEE KDEX workshop.

Mobasher, B., Cooley, R., and Srivastava, J. (1999). *Creating adaptive web sites trough usage based clustering of URLs*. Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange. IEEE Computer Society.

Mobasher, B. Dai, H. Luo, T. Nakagawa, M. (2001*). Effective personalization based on association rule discovery from web usage data*. in: Proc. of wIDM.

Molodtsov, D. (1999). *Soft set theory-first results*. Computers and Mathematics with Applications. 37, 19–31.

Ngu, D.S.W. and Wu, X. Sitehelper. (1997). *A localized agent that helps incremental exploration of the world wide web*. Proceeding of 6th International World Wide Web Conference. Santa Clara, CA: ACM Press.

Ng, W. Chan, C. (2001). *WHAT: A web hypertext associated trail mining system*. in: Proc. 9th IFIP 2.6 Working Conf. On Database Semantics.

Pal, S.K., Talwar, V. and Mitra, P. (2002). *Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions*. IEEE Transactions on neural network, 13 (5), , 1163 - 1177.

Pawlak, Z. (1982). *Rough sets*. International Journal of Computer and Information Science. 11, 341-356.

Pawlak, Z. (1991). *Rough sets: A theoretical aspect of reasoning about data*. Kluwer Academic Publisher.

Pawlak, Z. and Skowron, A. (2007). *Rudiments of rough sets*. Information Sciences. An International Journal. 177 (1), 3-27.

Pei, J. Han, J. Mortazaviasl, B. Zhu, H. (2000). *Mining access patterns efficiently from web logs.* in: Proc. PaciFc Asia Conf. on Knowledge Discovery and Data Mining, Kyoto, Japan.

Perkowitz, M. and Etzioni, O. (1998). *Adaptive Web Sites: Automatically Synthesizing Web Pages*. Proceedings of the 15th National Conference on Artificial Intelligence. Madison, WI: AAAI.

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

Schechter, S. Krishnan, M. Smith, M.D. (1998). *Using path profiles to predict HTTP requests*. Comput. Networks ISDN Systems 30 457–467.

Spiliopoulou, M. (1999). *Data mining for the web*. in: Principles of Data Mining and Knowledge Discovery, Second European Symp., PKDD'99pp. 588–589.

Srivastava, J. Cooley, R. Deshpande, M. Tan, P.-N. (2000). *Web usage mining: discovery and applications of usage patterns from web data*. in: SIGKDD Explorations, vol. 1 (2), pp. 12–23.

Steinbach , M. (2000). *A Comparison of Document Clustering Techniques*, KDD'2000, Technical report of University of Minnesota.

Wang, W. Zaiane, O.R. (2002). *Clustering web sessions by sequence alignment*. in: Proc. of DEXA.

Xiao, J. Zhang, Y. (2001). *Clustering of web users using session-based similarity measures*. in: Proc. of ICCNMC'01.

Xu, R. D. (2005). *Wunsch, Survey of clustering algorithms*. IEEE Transactions on Neural Networks 16 (3).

Yanchun, Z. Guandong, X. Xiaofang, Z. (2005). *A Latent Usage Approach for Clustering Web Transaction and Building User Profile.* Springer-Verlag Berlin Heidelberg, 31 - 42.

Yang, Q. Zhang, H.H. Li. T. (2001). *Mining web logs for prediction models in WWW caching and prefetching.* in: Proc. of ACM SIGKDD.

Zahn, C.T. (1971). *Graph-theoretical methods for detecting and describing gestalt structures*, IEEE Transactions on Computers C-20 68–86.

Zhao, Y. Karypis, G. (2005). *Hierarchical clustering algorithms for document datasets*, Data Mining and Knowledge Discovery 10 141–168.