Mining and Analysis of Clickstream Patterns

H. Hannah Inbarani and K. Thangavel

Abstract. The explosive growth of the web has drastically changed the way in which information is managed and accessed. The large-scale of web data sources and the wide availability of services over the internet have increased the need for effective web data mining techniques and mechanisms . A sophisticated method to organize the layout of the information and assist user navigation is therefore particularly important. In this work, we focus on web usage mining, applying data mining techniques to web server logs. Web usage mining is the non-trivial process of distinguishing implicit, previously unknown but potentially useful clickstream patterns that may exist in any collection of web access logs. The required abstraction can be generated by clustering the web access logs based on some sort of similarity measure. Clustering is done such that the web access logs within the same group or cluster are more similar than data points from different clusters. In this chapter, we propose a partitional algorithm namely Multi Pass Combined Standard Deviation(CSD) Means algorithm which automatically generates the optimum number of clusters from the web clickstream patterns. The quality of clusters obtained using these algorithms are compared using K-Means algorithm, Rough K-Means algorithm and model based algorithms ANTCLUST and ACCANTCLUST. The experimental analysis of mined clickstream patterns shows the effectiveness of the proposed algorithm.

Keywords: Clickstreams, Clustering, K-Means, Ant-Clustering, Rough K-Means.

1 Introduction

In the highly competitive world and with the broad use of the web in E-commerce, E-learning, and E-news, finding users' needs and providing useful information are the primary goals of web site owners. Therefore, analyzing clickstream patterns of web users becomes increasingly important[42]. This increase stems from the realization that added value for web site visitors is not gained merely through larger quantities of data on a site, but through easier access to the required information at the right time and in the most suitable form. With competitors being 'one-click away', the requirement for adding value to E-services on the web

H. Hannah Inbarani and K. Thangavel
Department of Computer Science, Periyar University, Salem-636 011
e-mail: hhinba@yahoo.co.in, indrktvelu@yahoo.com

A. Abraham et al. (Eds.): Foundations of Comput. Intel. Vol. 6, SCI 206, pp. 3–27.
springerlink.com

has become a necessity towards the creation of loyal visitors for a web site. This added value can be realized by focusing on specific individual needs and providing tailored products and services. Web currently constitutes one of the largest dynamic data repositories. So, it becomes necessary to extract useful knowledge from this raw and dynamic data by knowledge discovery called as web mining on the Internet.

Recently many researchers have proposed a new unifying area for all methods that apply data mining to web data, named web mining [9]. Web mining tools aim to extract knowledge from the web, rather than retrieving information. Commonly, web mining work is classified into the following three categories [5],[34] : web content mining, web usage mining and web structure mining. Web content mining is concerned with the extraction of useful knowledge from the content of web pages, by using data mining. Web structure mining is a new area, concerned with the application of data mining to the structure of the web graph.

Most of the researches in web usage mining techniques are used to discover user's browsing behaviour. This method processes the web clickstream data directly to find 'an interesting pattern' [41]. A click stream is then the sequence of page views that are accessed by the user. A user session is the click-stream of page views for a single user across the entire web [35],[31]. Originally, the aim of web usage mining has been to support the human decision making process and thus, the outcome of the process is typically a set of data models that reveal implicit knowledge about data items, like web pages, or products available at a particular web site. These models are evaluated and exploited by human experts, such as the market analyst who seeks business intelligence, or the site administrator who wants to optimize the structure of the site and enhance the browsing experience of visitors.

During the process of web usage mining, the rules and patterns in web log records are explored and analyzed mainly by means of the techniques relating to artificial intelligence, data mining, database theory and so on. A variety of machine learning methods have been used for pattern discovery in web usage mining. These methods represent the four approaches that most often appear in the data mining literature: clustering, classification, association discovery and sequential pattern discovery. Similar to most of the work in data mining, classification methods were the first to be applied to web usage mining tasks[9]. However, the difficulty of labeling large quantities of data for supervised learning has led to the adoption of unsupervised methods, especially clustering. The large majority of methods that have been used for pattern discovery from web data are clustering methods. Clustering aims to divide a data set into groups that are very different from each other and whose members are very similar to each other. The purpose of clustering users based on their web clickstreams in a particular web site is to find groups of users with similar interests and motivations for visiting that web site[9]. If the site is well designed there will be strong correlation between the similarity among user clickstreams and the similarity among the users' interests or intentions. Therefore, clustering of the former could be used to predict groupings for the latter.

One important constraint imposed by web usage mining is the choice of clustering method. In practical applications of clustering algorithms, several problems must be solved, including determination of the number of clusters and

evaluation of the quality of the partitions. In this research work, we explore the problem of clickstream clustering based on users navigation behavior. There are many methods and algorithms for clustering based on crisp [11], fuzzy [4], probabilistic [38] and possibilistic approaches [22].

Clusters can be hard or soft in nature. In conventional clustering, objects that are similar are allocated to the same cluster while objects that differ significantly are put in different clusters. These clusters are disjoint and are called hard clusters. In soft clustering, an object may be a member of two or more clusters. Soft clusters may have fuzzy or rough boundaries [24]. In fuzzy clustering, each object is characterized by partial membership whereas in rough clustering objects are characterized using the concept of a boundary region. A rough cluster is defined in a similar manner to a rough set. The lower approximation of a rough cluster contains objects that only belong to that cluster. The upper approximation of a rough cluster contains objects in the cluster which are also members of other clusters [38]. The advantage of using rough sets is that, unlike other techniques, rough set theory does not require any prior information about the data such as apriori probability in statistics and a membership function in fuzzy set theory. In this chapter, both hard and soft clustering techniques are used for clustering clickstream patterns.

In this chapter, we propose Multi Pass CSD Means algorithm [37] to cluster clickstream patterns. This intelligent algorithm automatically determines the expected number of clusters from the given clickstream patterns. A comparative analysis is made with other intelligent clustering algorithms such as ANTCLUST [28], ACCANTCLUST[18], K-Means algorithm[25] and Rough-K-Means algorithm[17]. Empirical results clearly show that the proposed Multi Pass CSD Means algorithm performs well and provides stable results compared with other clustering algorithms.

The rest of the chapter is organized as follows. In section 2, an overview of web usage mining process is described. In section 3, the work related to clickstream clustering is summarized. Section 4, discusses the work related to clickstream analysis. In section 5, the partitional and artificial intelligence based algorithms applied for clustering clickstream patterns are presented. In section 6, the experimental results and comparative analysis of the proposed algorithms are discussed. Section 7 concludes this chapter with directions for further research.

2 Web Usage Mining – An Overview

In general, web usage mining consists of three stages, namely data preprocessing, pattern discovery and pattern analysis. Web data collected in the first stage of data mining are usually diverse and voluminous. These data must be assembled into a consistent, integrated and comprehensive view, in order to be used for pattern discovery. As in most applications of data mining, data preprocessing involves removing and filtering redundant and irrelevant data, predicting and filling in missing values, removing noise, transforming and encoding data, as well as resolving any inconsistencies. The task of data transformation and encoding is particularly important for the success of data mining. In Web usage mining, this

stage includes the identification of users and user sessions, which are to be used as the basic building blocks for pattern discovery.

2.1 Data Preprocessing

2.1.1 Data Cleaning

The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining.

Data Filtering
Most web log records are irrelevant and require cleaning because they do not refer to pages clicked by visitors[47]. A user's request to view a particular page often results in several log entries for a single web page access since graphics and scripts are downloaded. Removing these irrelevant items can reduce the data that will be analysed and increase the analysis's speed. It also can decrease the irrelevant items' negative influence to the mining process.

Feature Selection
Log files usually contain nonessential information from the analytical point of view. Thus the first data pre-processing step is the selection of features. Moreover, reducing the number of features at this stage decreases the memory usage and improves performance[32]. It is also beneficial from the computational point of view, since log files contain thousands of megabytes of data. The final output of the pre-processing must be divided into sessions. The key attributes to build sessions are page ID, computer's IP address (or host) and page request time. These are the main features to work with in web usage mining. Other features are less relevant unless participating in some specific tasks.

Selamat and Omatu [32] proposed a neural network based method to classify web pages and use principle component analysis to select the most relevant features for the classification In [19], Quick Reduct and Variable Precision Rough Set (VPRS) Algorithms are proposed for feature selection from the web log file. These feature selection algorithms are used for selecting significant attributes (features) for describing a session which is suitable for pattern discovery phase.

2.1.2 User's Identification

User's identification is, to identify who access web site and which pages are accessed. If users have logged in their information, it is easy to identify them. In fact, there is lot of users who do not register their information. What's more is there are several users who access web sites through agent, several users use the same computer, firewall's existence, one user use different browsers, and so forth. All of these problems make this task greatly complicated and it is very difficult, to identify every unique user accurately.

2.1.3 Session Identification

For logs that span long periods of time, it is very likely that users will visit the web site more than once. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. The simplest method of achieving session is through a timeout, where if the time between page requests exceeds a certain time limit, it is assumed that the user is starting a new session. Many commercial products use 30 minutes as a default timeout.

2.1.4 Data Formatting

Data formatting [6] consists of mapping the number of valid URLs on a website to distinct indices. A user's clickstream consists of accesses originating from the same IP address within a predefined time period. Each URL in the site is assigned a unique number. Thus the pages visited by the users are encoded as binary attribute vectors.

2.1.5 Path Completion

Another critical step in data preprocessing is path completion. There are some reasons which result in path's incompletion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of URLs recorded in log may be less than the real one. This problem is referred to path completion, which will influence next steps' efficiency and accuracy if it is not solved properly.

2.2 Pattern Discovery

In the pattern discovery stage, machine learning and statistical methods are used to extract patterns of usage from the preprocessed web data[2]. A variety of machine learning methods have been used for pattern discovery in web usage mining. These methods represent the four approaches that most often appear in the data mining literature: clustering, classification, association discovery and sequential pattern discovery. Similar to most of the work in data mining, classification methods were the first to be applied to web usage mining tasks[9]. However, the difficulty of labeling large quantities of data for supervised learning has led to the adoption of unsupervised methods, especially clustering

2.3 Pattern Analysis

Pattern Analysis is the final stage of the whole web usage mining. The goal of this process is to eliminate the irrelative rules or patterns and to extract the interesting

rules or patterns from the output of the pattern discovery process. The output of web mining algorithms is often not in the suitable form for direct human consumption, and thus need to be transformed to a format that can be assimilated easily. There are two most common approaches for the patter analysis. One is to use the knowledge query mechanism such as SQL, while another is to construct multi-dimensional data cube before perform Online Analytical Processing (OLAP) operations. All these methods assume that the output of the previous phase has been structured.

3 Clickstream Clustering–Literature Review

Clustering aims to discover sensible organization of objects in a given dataset by identifying similarities as well as dissimilarities between objects. It classifies a mass of data, without any prior knowledge, into clusters which are clear in space partition outside and highly similar inside. In web usage mining, clustering algorithms can be used in two ways: usage clickstream clusters and page clickstream clusters[30].

3.1 Web User Clustering

An important research point in web usage mining is the clustering of web users based on their common properties. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in E-commerce applications or provide personalized web content to the users. On the other hand, clustering of pages will discover groups of pages having related content. Several researchers have applied data mining techniques to web server logs, attempting to unlock the usage patterns of web users hidden in the log files [13].

User profiling on the web consists of studying important characteristics of the web visitors. Due to the ease of movement from one portal to another, web users can be highly mobile. If a particular web site doesn't satisfy the needs of a user in a relatively short period of time, the user will quickly move on to another web site[1]. Therefore, it is very important to understand the needs and characteristics of web users. The clustering process is an important step in establishing user profiles. In web usage mining, clustering algorithms can be used in two ways: usage clusters and page clusters. Fu et al., [15] have demonstrated that web users can be clustered into meaningful groups, which help webmasters to better understand the users and therefore to provide more suitable, customized services. Perkowitz and Etzioni [29] proposed adaptive web sites that improve themselves by learning from user access patterns. The main mining subject is the web server logs.

In [45], Yan et al., the authors use the First Leader clustering algorithm to create groups of sessions. In [13], Estivill-Castro and et al., proposed a derivative form of the K-Means algorithms which uses the median instead of the barycentre to compute the position of the centre of the groups after each affectation of an object to a cluster. Nevertheless, the main problem still remains: the analyst has to

specify the number of expected clusters. To solve this problem, Heer and et al. proposed [18] to find automatically, the number of expected clusters by evaluating the stability of the partitions for different number of clusters. This method works well but is extremely time consuming. Thus, this method may not be used in the web mining context, where the data sets to explore contain a lot of sessions. Labroche and et al. described [16] a clustering algorithm called ANTCLUST that is inspired from the chemical recognition system of ants and that allows to find automatically the number of expected clusters. In[28], an enhanced version of ANTCLUST is used to handle the files of sessions extracted from the web server files. Lingras and West [22] provided a theoretical and experimental analysis of a modified K-Means clustering based on the properties of rough sets. In [16] , a common model-based clustering algorithm is used to result in clusters of web users' sessions.

The clustering methods presented in this chapter are restricted to generate user clusters. The terms user/session clusters is used interchangeably because the log file taken for the experiment consists of no date/time information and each user/session is identified only using user id and Page id.

3.2 Web Page Clustering

Clustering of pages will discover groups of pages having related content which could be useful for mass personalization and web site adaptation.

Flake et al., [14] use only link information to discover web communities (Groups of URLs). The web communities they discover however, merely reflect the viewpoint of web site developers. Mobasher et al., [26] proposed a technique of usage based clustering of URLs for the purpose of creating adaptive web site. They directly compute overlapping clusters of URL references based on their co-occurrence patterns across user transactions. But their URLs clustering is still based on frequent items and needs user session identification. Selamat and Omatu [32] proposed a neural network based method to classify web pages. Obviously, it is a content-based method, and it needs class-profile which contains the most regular words in each class. In [30], vector analysis based and fuzzy set theory based methods are used for the discovery of user clusters and page clusters.

4 Clickstream Analysis

On a web site, clickstream analysis is the process of collecting, analyzing, and reporting aggregate data about which pages visitors visit in what order - which are the result of the succession of mouse clicks of each visitor. Analyzing visitors' personal information gives us an idea of who might be a potential customer, but their current interests should be taken into consideration as well. Customers search for information about products they are interested in on the internet and the web server records every movement they make in a server log file. A visitor's changing interest is hidden in this file. Studies exploring customers' web page access patterns based on web log files are given in [35],[33],[40] and [46]. The volume of information stored in the web server log accumulates over time. The

more data processed, the more time needed to calculate results. In order to capture customers' current interests and provide instant service, on-line analysis must be performed. Discovering patterns of loyal customers' click streams stored in the log file is the province of off-line analysis, which does not require instant results. Web server log files are simple text files that are automatically generated every time someone accesses the web site. Every ''hit'' to the web site is logged in the form of one line of text. Information in the raw web log file format includes who the visitor was, where the visitor came from, and what he was doing on the site.

4.1 Common Web Log Format

A web server log file contains requests made to the web server, recorded in chronological order. The most popular log file formats are the Common Log Format (CLF) and the extended CLF.

A common log format file is created by the web server to keep track of the requests that occur on a web site. The format of a common log file [43] is shown in Table 1.

Table 1 Common Log file Format

remotehost	Remote hostname
Rfc931	The remote log name of the user.
Authuser	The username as which the user has authenticated himself
date	Date and time of the request.
request	The request line exactly as it came from the client
status	The HTTP status code returned to the client
Bytes	The content-length of the document transferred

The experiments in this chapter are conducted on the web access logs of www.microsoft.com which is available in UCI repository [http://www.ics.uci.edu/]. This web log consists of only page id to identify pages and user id to identify users.

The purpose of analyzing web logs is to understand the user's browsing behavior. Based on the result of perceived user's behavior, user's page searching time may be reduced by recommending pages customers may be interested in. The most important information revealed by analyzing customer's clickstream is the user's current interest.

Clustering analysis is widely used to establish object pro-files on the basis of objects' variables. Objects can be customers, web documents, web users, or facilities[19]. In this chapter clustering algorithms are applied on web clickstrems to analyze user access trends. On one hand, the profiles can be used for predicting the navigation behaviour of current users, thus aiding in web personalization. On the other hand, webmasters can improve the design and organization of websites based on the acquired profiles.

A sample web log file in the common log file format is shown in Table 2.

Table 2 Sample web log file

124.49.105.224 - - [29/Nov/2000:18:02:26 +0200] “GET /index.html HTTP/1.0” 200 1159
124.49.105.224 - - [29/Nov/2000:18:02:27 +0200] “GET /PtitLirmm.gif HTTP/1. 0” 200 1137
124.49.105.224 - - [29/Nov/2000:18:02:28 +0200] “GET /acceuil_fr.html HTTP/1. 0” 200 1150
124.49.105.224 - - [29/Nov/2000:18:02:30 +0200] “GET /venir/venir.html HTTP/1.0” 200 1141
124.49.105.225 - - [29/Oct/2000:18:03:07 +0200] “GET /index.html HTTP/1.0” 200 1051
124.49.105.225 - - [16/Oct/2000:20:34:32 +0100] “GET /formation.html HTTP/1.0” 200 -
124.49.105.225 - - [31/Oct/2000:01:17:40 +0200] “GET /formation.html#d HTTP/1.0” 304 -
124.49.105.225 - - [31/Oct/2000:01:17:42 +0200] “GET /theses2000.html HTTP/1.0” 304 -
124.49.105.56 - - [22/Nov/2000:11:06:11 +0200] “GET /lirmm/bili/ HTTP/1.0" 200 4280
124.49.105.56 - - [22/Nov/2000:11:06:12 +0200] “GET /lirmm/rev_fr.html HTTP/1.0" 200 -
124.49.105.56 - - [07/Dec/2000:11:44:15 +0200] “GET /ress/ressources.html HTTP/1.0" 200

5 Related Work

One of the major issues in web log mining is to group all the users’ page requests so to clearly identify the paths that users followed during navigation through the web site [1].

There are many methods applied in clustering analysis, such as hierarchical clustering, partition-based clustering, density-based clustering, and artificial intelligence based clustering. In this chapter, artificial intelligence based clustering algorithms and partition-based clustering algorithms are studied for clustering clickstream patterns. The traditional clustering algorithms such as K-Means require users to provide the correct (actual) number of clusters in a given pattern set at the beginning. However, in many applications, a priori knowledge of the actual number of clusters is unavailable, and the optimal number of clusters cannot be easily and intuitively estimated beforehand. If the number of clusters estimated is larger than the actual number of clusters, one good compact cluster in nature is divided into more compact clusters with inappropriate separations; in contrast, if the estimated number of clusters is smaller, two or more compact clusters in nature must be grouped into one loose cluster. Thus, how to determine

the optimal number of clusters in a given pattern set is an important problem in cluster analysis. It is very difficult to identify the optimal number of clusters on the access sessions from the clustering results in an unsupervised way because this number determines how many representative navigation patterns will be extracted from user access sessions, and how many user profiles are supposed to be constructed next. The optimal number means that the partition of user access sessions can best reflect the distribution of sessions, and can also be validated by user's inspection [8]. All the existing algorithms discussed in this chapter except K-Means and Rough K-Means estimates optimal number of clusters from the given web log data set.

5.1 Artificial Intelligence Based Clustering Algorithms

Applying ant colony system in clustering analysis is still a very novel research area. Ant Colony Algorithm (ACA) is a meta-heuristic approach successfully applied to solve hard combinatorial optimization problems [48]. It is also feasible for clustering analysis in data mining. Many researchers use ant algorithms for clustering analysis and the result is better than other heuristic methods. Unlike traditional clustering methods, ACA is an intelligent approach, which is successfully used in discrete optimization. The ant-based clustering algorithm is inspired by the behavior of ant colonies in clustering their corpses and sorting their larvae and automatically finds the number of clusters .

Algorithm: ANTCLUST

1) Intitialisation of the ants.
2) $Genetic_i \leftarrow X_i$, i^{th} session of the data set.
3) $Label_i \leftarrow 0$.
4) $Template_i$ is initialized.
5) $M_i \leftarrow 0$, $M_i^+ \leftarrow 0$, $A_i \leftarrow 0$.
6) Simulate N_{iter} iterations during which two ants that are randomly chosen meet.
7) Delete nests less than P * N (P<1) ants
8) Re-assign each ant having no more nest to the nest of the most similar ant found that have a nest.

Fig. 1 ANTCLUST

5.1.1 Antclust

This algorithm is based on real ants' collective behaviour, namely the construction of colonial odour and its use for determining the ants nest membership. Every day, the real ants have to solve a crucial recognition problem when they meet: they have to decide whether they belong to the same nest or not, in order to guaranty

the survival of the nest. This phenomenon is called "colonial closure" [25]. It manly relies on continuous exchanges and updates of chemical cues on the ant's cuticle and in their post-pharyngeal gland, determining, as an identity card, their belonging to the nest. Thus, each ant has its own view of its colony odor at a given time, and updates it continuously. By this way, an ant preserves its nest from being attacked by predators or parasites and reinforces its integration in nest. The gathering of ants in a finite number of nests where nest-mates are more similar to each other than the ants of other colonies provides a cluster of the set of objects. In this chapter, ANTCLUST[26] and accelerated ant clustering algorithm [16] are used for comparative analysis.

5.1.2 ACCANTCLUST

In ANTCLUST algorithm, when meeting between two ants is simulated, if the meeting is between two ants with no nest, and if they accept each other, these two ants are placed in a new nest. If they do not accept each other, no nest is created. In ACCANTCLUST, if the ants do not accept each other, two new nests are created and the ants are placed in two different nests.

Algorithm: ACCANTCLUST

Initialization of the ants

1) ∀ants(sessions) i ε [1,n]
2) Genome i ⟵ X_i, i th session vector of the data set
3) $Label_i$ ⟵ 0
4) Template is initialized
5) M_i ⟵ 0 M_i^+ ⟵ 0 A_i ⟵ 0
6) N_{biter} ⟵ 50 * n
7) Simulate Nbiter iterations during which two randomly chosen ants(sessions) meet.
   - i) If Label of i^{th} and j^{th} ant are zero and
     if the acceptance is true,
     Place them in the same nest
     Else
     Create two different nests and place them separately
   - ii) If Label of i^{th} ant is zero and j^{th} ant is not zero and
     If the acceptance is true
     Adding an ant with no label to an existing nest:
   - iii) Positive" meeting between two nestmates:
     Increase the values of Parameters M_i and M_i^+
   - iv) Negative" meeting between two nestmates:
     The worst integrated ant is removed from the nest and its label is set to zero.
   - v) Meeting between two ants of different nests:
     The ant x with the lowest M_x changes its nest and belongs now to the nest of the encountered ant.

Fig. 2 Accelerated Ant Colony Algorithm

5.2 Partitional Clustering Algorithms

In partitional clustering algorithms, each cluster can be represented by its center; thus, the solution of the partitional clustering algorithms can be represented by a set of clusters or a set of cluster centers [39]. Partitional algorithms construct a partition of a database D of n objects into a set of K clusters, where K is an input parameter for these algorithms. To set the value of K, some domain knowledge is required which unfortunately is not available in many applications such as clustering of web clickstream patterns.

5.2.1 K-Means Algorithm

The K-Means algorithm is the most well-known partitional clustering method due to its easy implementation and rapid convergence. This algorithm iteratively updates the solution in a deterministic manner such that their results are heavily influenced by the choice of initial solution. It is the simplest and most commonly used algorithm that employs a squared error criterion [25]. Provided with a set of n numeric objects and an integer number K($K \leq n$), it calculates a partition of patterns in K clusters. This process takes place in an iterative manner starting from a random initial partition and keeping on searching for a partition of n that minimizes the within groups sum of squared errors.

Algorithm: K-MEANS

1) Choose K initial cluster centres from the set of sessions $Z_1, Z_2 \ldots Z_k$.
2) At the k-th iterative step, distribute the user sessions $\{X_n\}$ among the K clusters using the relation
   $X_n \in C_j(k)$ if $\| X_n - Z_j(k) \| < \| X_n - Zi(k) \|$
   for all i=1, 2, … ,K; $i \neq j$; where $C_j(k)$ denotes the set of user sessions whose cluster centre is $Z_j(k)$.
3) Compute the new cluster centres $Z_j(k+1)$, j=1, 2…K such that the sum of the squared distances from all points in $C_j(k)$ to the new cluster centre is minimized. The measure which minimizes this is simply the sample mean of $C_j(k)$. Therefore, the new cluster centre is given by
   $$Z(k+1) = \sum_{x \in Cj(k)} X_n \qquad j = 1,2, \ldots K$$
   where N_j is the number of samples in $C_j(k)$
4) If $Z_j(k+1) = Z_j(k)$ for j = 1,2,…K then the algorithm has converged and the procedure is terminated.
5) Otherwise go to step 2.

Fig. 3 K-Means algorithm

5.2.2 Rough Clustering

In rough clustering each cluster has two approximations, a lower and an upper approximation. The lower approximation is a subset of the upper approximation. The members of the lower approximation belong certainly to the cluster, therefore they cannot belong to any other cluster. The data objects in an upper approximation may belong to the cluster. Since their membership is uncertain they must be a member of an upper approximation of at least another cluster[17. There is a crucial difference to fuzzy set theory where we have overlapping clusters too: in fuzzy set theory an object can belong to many sets; in rough sets the memberships to two or more sets indicate that there is information missing to determine the actual membership to one and only one cluster.

Algorithm: Rough K-Means

---

1) Assign the session vectors to the approximations.
   (i) For a given session vector X_n ,determine its closest mean m_h

$$d(X_l, Z_h) = \min_{n,k} d(X_n, Z_k) \Rightarrow X_l \in \overline{C_k} \wedge X_l \in \underline{C_k}$$

   Assign X_n to the upper approximation of the cluster h, $X_n \in \overline{C_h}$

   (ii) Determine the means m_t that are also close to X_n they are not farther away from X_n than d (X_n, Z_h) + ε where ε is a given threshold:

$$d(X_l, Z_h) = \min_{n,k} d(X_n, Z_k) \Rightarrow X_l \in \overline{C_k} \wedge X_l \in \underline{C_k}$$

   If $T \neq \emptyset$ (X_n is also close to at least one other mean Z_t besides Z_h)
   Then $X_n \in \overline{C_l} \ \forall \ t \in T$
   Else $X_n \in \underline{C_h}$

2) Compute new mean for each cluster C_i using the following equation

$$Z_k = W_l \sum_{X_n \in \underline{C_k}} \frac{\overline{X_n}}{|\underline{C_k}|} + W_b \sum_{X_n \in \underline{C_k^B}} \frac{\overline{X_n}}{|C_k^B|} \quad \text{for } C_k^B \neq 0$$

$$W_l \sum_{X_n \in \underline{C_k}} \frac{\overline{X_n}}{|\underline{C_k}|} \quad \text{otherwise}$$

   with $W_l + W_u = 1$.

Fig. 4 Rough K-Means algorithm

As defined in [15], for stable results, the parameter values are taken as W_l =0.7, W_u = 0.3, and threshold is taken as 0.04 since the session matrix constructed from the web log file is a sparse matrix.

5.2.3 Multi Pass CSD Means Clustering Algorithm

This section proposes Multi Pass CSD Means algorithm for clustering web clickstreams. In this algorithm, web users are grouped into the corresponding available clusters when they are similar. On the other hand, the dissimilar or irrelevant sessions are placed into the new cluster created. This shows, the algorithm remains adaptive in response to significant events by existing clusters and yet remains stable to irrelevant events by creating new cluster[37]. It retains the centroid of the available clusters. This process is repeated till all the clickstream patterns in the web log data set residing are considered. The multi-pass CSD-means algorithm is described as follows:

Algorithm: Multi Pass CSD Means

---

1) Assign the first session vector as the initial centroid for the first cluster
2) Assign each session vector to the available clusters with nearest center or Create and assign to the new cluster

   i) For each session vector X_n
      Check class-set value for X_n in the available clusters
      Class-set = false;
   ii) For each cluster k do
      find the combined standard deviation of σ_{csd} of X_n and Z_k
      And the session with maximum standard deviation σ_{max} of X_n and Z_k
      if there exist σ_{max} which is less than σ_{csd} then it can be set to some classes break;
      if (class-set = true) then
         find the minimum distance and place it in the corresponding cluster such as K-Means algorithm
3) if current centroid and old centroid are not very close repeat the process
4) find minimum distance to a center and place the session into the corresponding cluster

Stop.

Fig. 5 Multi Pass CSD Means

6 Experimental Results

6.1 Data source

The data source of web usage mining is web log files, from which we can realize users' clickstream patterns by web usage mining. For a web site, user access information is generally gathered automatically by web server and recorded in the server logs. There are four different log files, the access log, agent log, error log, and referrer log. These log files are text files, and their sizes depend on the traffic at a particular site. Recorded in these files is the volume of activity at each page on a web site, the type of browser used to access each page, any errors that users may have experienced downloading pages from the web site, and where users were referred from when they accessed pages at the web site.

For the purpose of evaluating the performance and the effectiveness of the intelligent clustering algorithms, experiments were conducted with preprocessed web access logs of www.microsoft.com which is available in UCI repository [http://www.ics.uci.edu/]. This log file records the use of www.microsoft.com by 5000 anonymous, randomly-selected users who have visited the web site in a one week timeframe in February 1998 with an average of 5.7 page views per user. The file contains no personally identifiable information. This data set includes visits which are recorded are recorded in time order and no pre-processing is required since data set was given in sessions. The 294 web pages are identified by their title (e.g. "NetShow for PowerPoint") and URL (e.g. "/stream"). These algorithms are applied only for testing instances available in UCI repository by taking only 100 web pages and 5000 users. The five data sets with sizes of 1000, 2000,3000, 4000,5000 users were extracted from the log file of Microsoft web site.

6.2 Data Preparation

As the web log file of Microsoft web site available in UCI repository is a preprocessed one, the only preprocessing step needed is data formatting. The only fields available in this log file are user id and id of page visited . Since there is no date and time information in the given web log, all the pages visited by the user are considered as a single session. Using this information session (user access) matrix is constructed.

6.3 Web Traffic Patterns

Fig 6. shows the web traffic patterns of users who have visited the Microsoft web site in a one week timeframe in February 1998 and the average number of page views is per user is 5.7.

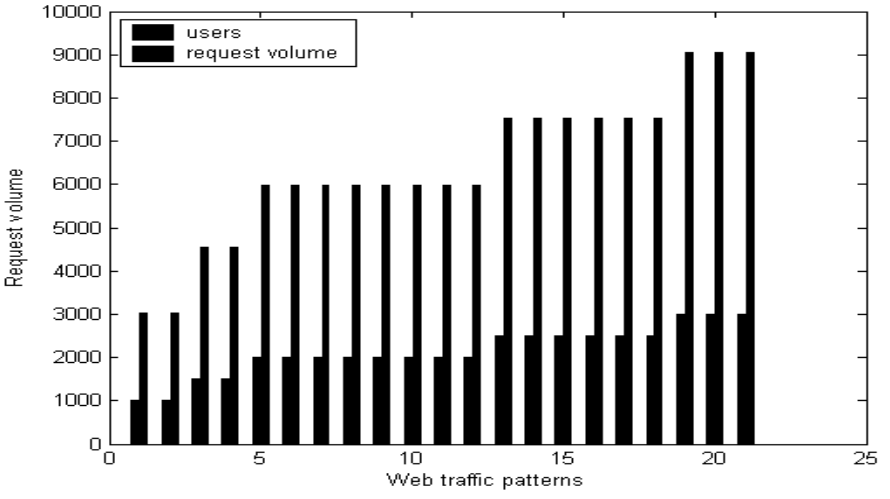

Fig. 6 Web Traffic Patterns of Microsoft web site

6.4 Results of the Proposed Algorithm

6.4.1 User Profiles

User profiling on the web consists of studying important characteristics of the web visitors. The clustering process is an important step in establishing user profiles.

Some of the discovered user profiles obtained using Multi Pass CSD Means algorithm are summarized in Table 3. These user profiles are obtained by clustering clickstreams only 100 users from the web log file.

The extracted user profiles reveal that most of the users who have visited the web site are interested in various Microsoft software's , some of the users are interested in internet tools and others interested in network related softwares. Single users in clusters show that those users are interested in country specific web pages. On one hand, the profiles can be used for predicting the navigation behaviour of current users, thus aiding in web personalization. On the other hand, webmasters can improve the design and organization of web sites based on the acquired profiles.

6.4.2 User and Page Distribution of Clusters

It is also very useful to have a view about the contents of each cluster, since a deeper knowledge for the inside of each cluster can draw useful and meaningful inferences for the users' navigation behavior.

We also extracted a brief summary for each cluster by computing a mean vector for each cluster and extracted the corresponding web pages accessed in each cluster. The results of applying Multi Pass K-Means algorithm for 5000 users

Table 3 Sample User Profiles obtained using Multi Pass CSD-Means algorithm

Cluster	Size	URLs/Profile	Profile Descriptions
1	19	/mexico /homeessentials /kids /msp /support /vstudio /publisher /activex /products /ntworkstation /proxy /kb /windowssupport /catlog /teammanager /officefreestuff /workshop /msdn /iis	Mexico, Microsoft Home Essentials, MSHome Kids Stuff, Microsoft Solution Providers, Support Desktop, Visual Studio, MS Publisher, ActiveX Technology Development, Products, Windows NT Workstation, MS Proxy Server, Knowledge Base, Windows95 Support, Product Catalog, MS TeamManager, Office Free Stuff, Developer Workshop, Developer Network, Internet Information Server
2	1	/Taiwan	Taiwan
3	8	/sbnmember /ie_intl /intdev /netmeeting /activex /java /workshop /msdn	SiteBuilder Network Membership, Internet Development, International IE content, NetMeeting, ActiveX Technology Development, Java Strategy and Info, Developer Workshop, Developer Network
4	8	/windowssupport /kb /supportnet /mspress /products /iesupport /security /support	Windows95 Support, Knowledge Base, Support Network Program Information , Products , Microsoft Press, IE Support, Internet Security Framework, Support Desktop
5	2	/intdev /workshop	Internet Development, Developer Workshop
6	2	/regwiz /support	Regwiz, Support Desktop
7	1	/msp	Microsoft Solution Providers
8	4	/education /support /truetype /catalog	MS in Education, Support Desktop, Typography Site, Product Catalog
9	1	/windows support	Windows95 Support
10	4	/promo /automap /organizations /frontpage	Promo, N. American Automap, Corporate Desktop Evaluation, FrontPage
11	4	/referral /frontpage /uk /msdn	SP Referral (ART), FrontPage, UK, Developer Network
12	1	/ie_intl	International IE content
13	1	/workssupport	Works Support
14	4	/office reference /organizations /regwiz /officefreestuff	Office Reference, Corporate Desktop Evaluation, regwiz, Office Free Stuff
15	2	/Canada /referral	Canada, SP Referral (ART)
16	3	/ntwkssupport /promo /ntwkssupport	NT Workstation Support, promo, NT Workstation Support

taken from web log file is summarized below. The optimum number of clusters generated by the proposed algorithm for 5000 users is 48. For 48 clusters, we computed the mean vector and extracted the corresponding web pages accessed in each cluster based on the values of its mean vector. Table 4 gives a specific description on each of the obtained, the number of users in each cluster and the percentage of users in the cluster. It can be seen from the table that the forty eighth cluster accounts for the largest proportion of users.

Table 5 gives a specific description on each of the obtained number of pages distributed in each user cluster and the percentage of pages visited by the users in the cluster. It can be observed from the table that the cluster four encounters larger number of web pages.

More specifically, Fig. 7 depicts the percentage frequency of requested users and web page categories observed in each cluster for 5000 users by applying proposed algorithm to help in understanding users' navigation behavior for the web clickstream patterns.

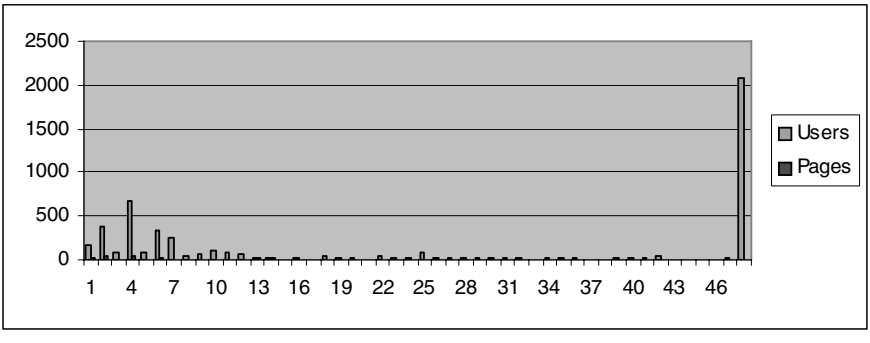

Fig. 7 User and Page request volume of clusters

6.5 Comparative Analysis

6.5.1 Cluster Validation

Cluster validation refers to procedures that evaluate the clustering results in a quantitative and objective function. The validation index is a single real value that can describe the quality of a complete cluster partition Some kinds of validity indices are usually adopted to measure the adequacy of a structure recovered through cluster analysis In fact, if cluster analysis is to make a significant contribution to engineering applications, much more attention must be paid to cluster validity issues that are concerned with determining the optimal number of clusters and checking the quality of clustering results. Many different cluster validity measures have been proposed such as the Dunn's separation measure [12], the Bezdek's partition coefficient [3], the Xie-Beni's separation measure [44], Davies-Bouldin's measure [8], *etc*. In this chapter , the popular validity measures

Table 4 User distribution of clusters

Cluster label	Number of Users	Percentage Of Users
1	167	0.3
2	376	0.39
3	79	0.07
4	675	0.43
5	81	0.04
6	343	0.24
7	244	0.07
8	36	0.02
9	72	0.09
10	97	0.08
11	88	0.08
12	61	0.08
13	23	0.19
14	17	0.15
15	9	0.02
16	14	0.01
17	9	0.02
18	36	0.01
19	25	0.02
20	23	0.06
21	8	0.02
22	33	0.02
23	27	0.01
24	13	0.02
25	74	0.05
26	14	0.02
27	11	0.03
28	21	0.01
29	31	0.06
30	13	0.02
31	16	0.1
32	17	0.01
33	6	0.01
34	11	0.02
35	13	0.07
36	12	0.02
37	9	0.03
38	4	0.01
39	13	0.01
40	12	0.01
41	13	0.02
42	33	0.01
43	4	0.01
44	1	0.01
45	2	0.01
46	5	0.01
47	25	0.01
48	2084	0.01

Table 5 Page distribution of clusters

Cluster label	Number of Pages	Percentage of Pages
1	30	0.3
2	39	0.39
3	7	0.07
4	43	0.43
5	4	0.04
6	24	0.24
7	7	0.07
8	2	0.02
9	9	0.09
10	8	0.08
11	8	0.08
12	8	0.08
13	19	0.19
14	15	0.15
15	2	0.02
16	1	0.01
17	2	0.02
18	1	0.01
19	2	0.02
20	6	0.06
21	2	0.02
22	2	0.02
23	1	0.01
24	2	0.02
25	5	0.05
26	2	0.02
27	3	0.03
28	1	0.01
29	6	0.06
30	2	0.02
31	10	0.1
32	1	0.01
33	1	0.01
34	2	0.02
35	7	0.07
36	2	0.02
37	3	0.03
38	1	0.01
39	1	0.01
40	1	0.01
41	2	0.02
42	1	0.01
43	1	0.01
44	1	0.01
45	1	0.01
46	1	0.01
47	1	0.01
48	1	0.01

such as the Davies-Bouldin's measure, Xie-Beni's separation measure and Dunn's index are used to evaluate the web clickstream clusters and the results are shown in Table 6. For the evaluation of Rough-K-means algorithm , rough version of Davies-Bouldin's measure and Dunn's index [36] is used.

Table 6 Cluster Validity Results

No.of Users	Page Request volume	Clustering Algorithm	No.of Clusters (Input)	No.of Clusters Generated	Davies Bouldin Index	Dunn's Index	Xie-Beni's Index
1000	3033	ANTCLUST	-	57	0.41	0.98	0.71
		ACCANTCLUST	-	41	0.39	1.13	0.68
		K-Means	5	-	0.307	1.289	0.549
		Rough K-Means	5	-	0.373	1.31	0.55
		Multi Pass CSD	-	37	0.32	1.33	0.67
2000	5992	ANTCLUST	-	62	0.403	1.296	0.59
		ACCANTCLUST	-	47	0.437	0.98	0.64
		K-Means	7	-	0.3	1.39	0.613
		Rough K-Means	7	-	0.362	1.01	0.62
		Multi Pass CSD	-	40	0.311	1.37	0.58
3000	9062	ANTCLUST	-	71	0.423	0.912	0.59
		ACCANTCLUST	-	52	0.414	1.10	0.54
		K-Means	9	-	0.519	1.117	0.642
		Rough K-Means	9	--	0.46	1.05	0.69
		Multi Pass CSD	-	41	0.324	1.279	0.578
4000	12066	ANTCLUST	-	74	0.517	0.928	0.73
		ACCANTCLUST	-	61	0.501	1.11	0.58
		K-Means	11	-	0.579	0.934	0.623
		Rough K-Means	11	-	0.410	1.19	0.69
		Multi Pass CSD	-	43	0.338	1.24	0.52
5000	14960	ANTCLUST	-	82	0.44	0.98	0.77
		ACCANTCLUST	-	63	0.41	0.95	0.78
		K-Means	13	-	0.567	0.99	0.623
		Rough K -Means	13	-	0.48	1.01	0.641
		Multi Pass CSD	-	45	0.38	1.227	0.593

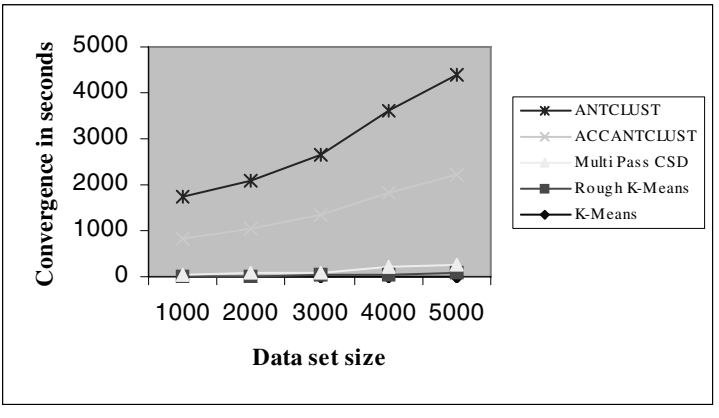

Fig. 8 CPU Time taken by Clustering algorithms

If a data set contains well-separated clusters, the distances among the clusters are usually large and the diameters of the clusters are expected to be small. Therefore larger value of Dunn's index (DI) means better cluster configuration. The Davies - Bouldin(DB) index is based on similarity measure of clusters whose bases are the dispersion measure of a cluster and the cluster dissimilarity measure. The

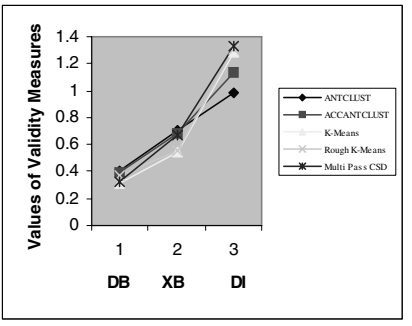

Fig. 9 Data set size = 1000

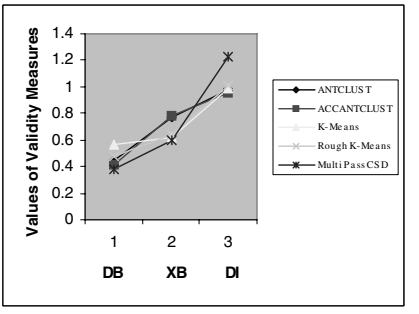

Fig. 10 Data set size = 2000

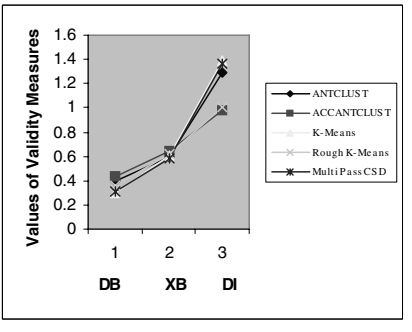

Fig. 11 Data set size = 3000

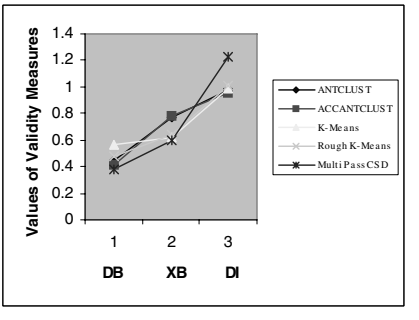

Fig. 12 Data set size = 4000

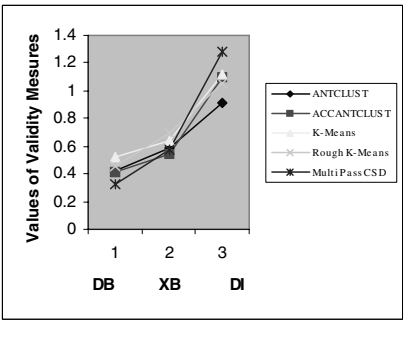

Fig. 13 Data set size = 5000

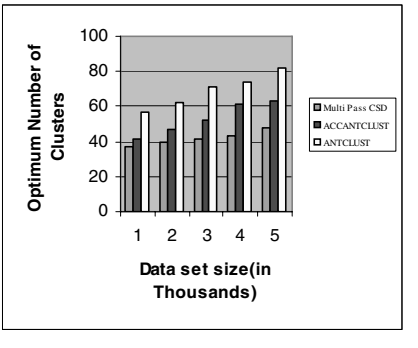

Fig. 14 Comparison of Algorithms

Davies–Boludin index measures the average of similarity between each cluster and its most similar one. As the clusters have to be compact and separated the lower Davies–Bouldin index means better cluster configuration. Xie-Beni (XB) validity index which measures the compactness and separation of clusters. Thus a smaller value of the index reflects that the clusters have greater separation from each other and are more compact.

Table 6 shows the effectiveness of the proposed algorithm. For each set of users taken, the value of Davies–Bouldin index for the proposed algorithm is lower than the value of Davies–Bouldin index for other algorithms. The value of Dunn's index for each set of users taken for the proposed algorithm is larger than the Dunn's index value of other algorithms. Figures 9 to 13 show the values of DB, DI and XB for the algorithms proposed for various sizes of the data sets. The optimum number of clusters generated by various algorithms is depicted in Fig 14.

In K-Means and Rough K-Means algorithm, the accuracy of the clusters obtained depends upon the number of clusters selected and the initial value of seed points. From the table 6, it can be observed that the proposed algorithm shows the accurate results since it estimates the number of clusters automatically from the given data set. Though ANTCLUST and ACCANTCLUST estimate the optimum number of clusters, these two algorithms take large amount of CPU time and the memory consumption is also very high. So it is not suitable for very large data sets. Fig 8 shows the CPU time taken for the covergence of clustering algorithms for each data set . We performed experiments on 2.4 GHz, 512MB RAM, Pentium-IV machine running on Microsoft Windows XP .

7 Conclusion

In this chapter, we investigated the proposed Multi-Pass CSD-Means algorithm for clustering web clickstream patterns from web access logs of www.microsoft.com which is available in UCI repository. A meaningful contribution for the clustering analysis of web logs can be made by using these algorithms. This algorithm estimates the optimum number of clusters automatically and is capable of manipulating efficiently very large data sets. A comparative analysis is made with K-Means algorithm, Rough K-Means and Model based algorithms ANTCLUST and ACCANTCLUST. The study of validity measure shows the effectiveness of the proposed algorithm for mining web clickstream patterns. Experimental results also show the stability and accuracy of the proposed algorithm. Though ANTCLUST and ACCANTCLUST estimates the number of clusters from the data set, experimental results clearly show that they are not stable and these algorithms take enormous amount of time for the convergence of clusters. Further research may be extended to computing optimum number of clusters for soft clusters.

References

1. Abraham, A.: Natural Computation for Business Intelligence from Web Usage Mining. In: Proceedings of the Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2005) (2005)
2. Baumgarten, M., Bchner, A.G., Anand, S.S., Mulvenna, M.D., Hughes, J.G.: Navigation Pattern Discovery from Internet Data. In: Masand, B., Spiliopoulou, M. (eds.) WebKDD 1999. LNCS, vol. 1836. Springer, Heidelberg (2000)

3. Bezdek, J.C.: Numerical Taxonomy with Fuzzy Sets. J. Math. Biol. 1, 57–71 (1974)
4. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York (1981)
5. Cooley, R.: Web Usage Mining: Discovery and Application of Interesting Patterns from web data, Ph.D. Thesis, University of Minnesota (2000)
6. Cooley, R., Mobasher, B., Srivastava, J.: Data Preparation for Mining World Wide Web Browsing Patterns. J. Knowledge and Information Systems 1(1), 5–32 (1999)
7. Cooley, R., Srivastava, J., Mobasher, B.: Web Mining: Information and Pattern Discovery on the World Wide Web. In: Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 1997), pp. 558–567 (1997b)
8. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Analysis and Machine Intelligence 1(4), 224–227 (1979)
9. Pierrakos, D., Paliouras, G.O., Papatheodorou, C., Spyropoulos, C.D.: Web Usage Mining as a Tool for Personalization: A Survey. User Modeling and User-Adapted Interaction 13, 311–372 (2003)
10. Dubes, R., Jain, A.K.: Validity studies in clustering methodologies. Pattern Recognition 11(1), 235–253 (1979)
11. Duda, R., Hart, P.: Pattern Classification and Scene Analysis. Wiley Interscience, New York (1973)
12. Dunn, J.C.: A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters. Journal Cybern. 3(3), 32–57 (1973)
13. Estivill-Castro, V., Yang, J.: Fast and robust general purpose clustering algorithms. In: Pacific Rim International Conference on Artificial intelligence, pp. 208–218 (1979)
14. Flake, G.W., Lawrence, S., Lee Giles, C., Coetzee, F.M.: Self- organization and identification of Web communities. IEEE Computer 35(3), 66–71 (2002)
15. Fu, Y., Sandhu, K., Shi, M.: Clustering of web users based on access patterns. In: Masand, B., Spiliopoulou, M. (eds.) WebKDD 1999. LNCS (LNAI), vol. 1836, pp. 21–38. Springer, Heidelberg (2000)
16. Pallis, G., Angelis, L., Vakali, A.: Validation and interpretation of Web users' sessions clusters. In: Information Processing and Management (2006)
17. Peters, G.: Some refinements of rough k-means clustering. Pattern Recognition 39, 1481–1491 (2006)
18. Hannah Inbarani, H., Thangavel, K.: Clickstream Intelligent Clustering using Accelerated Ant Colony Algorithm. In: Advanced Computing and Communications, 2006. ADCOM 2006. International Conference I, pp. 129–134 (2006)
19. Hannah Inbarani, H., Thangavel, K., Pethalakshmi, A.: Rough Set Based Feature Selection for Web Usage Mining. In: Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007), pp. 33–38 (2007) ISBN:0-7695-3050-8
20. Heer, J., Chi, E.: Mining the structure of user activity using cluster stability. In: Proceedings of the Workshop on Web Analytics, SIAM Conference on Data Mining, Arlington, VA (April 2002)
21. Chang, H.-J., Hung, L.-P., Ho, C.-L.: An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis. Expert Systems with Applications 32, 753–764 (2007)
22. Krishnapuram, R., Keller, J.: A possibilistic approach to clustering. IEEE Trans. Fuzzy Syst. 1(2), 98–110 (1993)

23. Kuo, R.J., Wang, H.S., Hu, T.-L., Chou, S.H.: Application of Ant K-Means on Clustering Analysis. Computers and Mathematics with Applications 50, 1709–1724 (2005)
24. Lingras, P., West, C.: Interval Set Clustering of Web Users with Rough K-means. Journal of Intelligent Information Systems (2002)
25. McQueen, J.: Some methods for classification and analysis of multivariate observations. In: Le Cam, L.M., Newman, J. (eds.) Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)
26. Mobasher, B., Cooley, R., Srivastava, J.: Creating adaptive web sites through usage-based clustering of URLs. In: Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX) (1999)
27. Labroche, N., Monmarche, N., Venturini, G.: A new clustering algorithm based on the chemical recognition system of ants. In: Proceedings of 15th European Conference on Artificial Intelligence (ECAI 2002), Lyon FRANCE, pp. 345–349 (2002)
28. Labroche, N., Monmarche, N., Venturini, G.: Web session clustering with artificial ants colonies'. In: Proc. of WWW 2003, May 20-24 (2003)
29. Perkowitz, M., Etzioni O.: Adaptive sites: automatically learning from user access patterns. In: Proceedings of WWW6 (1997), http://www.scope.gmd.de/info/www6/posters/722/index.html
30. Song, Q., Shepperd, M.: Mining web browsing patterns for E-commerce. Computers in Industry 57, 622–630 (2006)
31. Bucklin, R.E., Lattin, J.M., Ansari, A., Gupta, S., Bell, D., Coupey, E., Little, J.D.C., Mela, C., Montgomery, A., Steckel, J.: Choice And the Internet: From Clickstream to Research Stream. Marketing Letters 13(3), 245–258 (2002)
32. Selamat, A., Sigeru, O.: Web page feature selection and classification using neural networks. Information Sciences 158, 69–88 (2004)
33. Song, A.-B., Zhao, M.-X., Liang, Z.-P., Dong, Y.-S., Luo, J.-Z.: Discovering user profiles for Web personalization recommendation. Journal of Computer Science and Technology 19(3), 320–328 (2004)
34. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.T.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations 1(2), 2–23 (2000)
35. Kumar De, S., Radha Krishna, P.: Clustering web transactions using rough approximation. Fuzzy Sets and Systems 148, 131–138 (2004)
36. Mitra, S.: Rough-Fuzzy Collaborative Clustering. IEEE Transactions on Systems, Man and Cybernetics 36(4) (2006)
37. Thangavel, K., Ashok Kumar, D.: Pattern Clustering using Neural Network, Vision 2020: The Strategic role of Operational Research. Allied Publishers PVT LTD, New Delhi, pp. 662–679 (2006)
38. Titterington, D., Smith, A., Makov, U.: Statistical analysis of finite mixture distributions. John Wiley and Sons, Chichester (1985)
39. Voges, K.E., Pope, N.K.L., Brown, M.R.: Cluster analysis of marketing data examining online shopping orientation: a comparison of k-means and rough clustering approaches. In: Abbass, H.A., Sarker, R.A., Newton, C.S. (eds.) Heuristics and Optimization for Knowledge Discovery, pp. 207–224. Idea Group Publishing, Hershey (2002)
40. Wang, X., Abraham, A., Smith, K.: Intelligent web traffic mining and analysis. Journal of Network and Computer Applications 28(2), 147–165 (2005)

41. WangBin, Liuzhijing: Web Mining Research. In: Proceedings of the Fifth nternational Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2003) (2003)
42. Xing, W., Ghorbani, A.: Weighted PageRank Algorithm. In: Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR 2004) (2004)
43. W.W.W. Consortium. The Common Log file Format (1995), http://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format
44. Xie, X.L., Beni, G.: A Validity Measure for fuzzy Clustering. IEEE Trans. on Pattern Analysis and MachineIntelligence 13(8), 841–847 (1991)
45. Yan, T.W., Jacobsen, M., Garcia-Molina, H., Dayal, U.: From user access patterns to dynamic hypertext linking. In: Proceedings of 5th WWW, pp. 1007–1014 (1996)
46. Zhang, X., Gong, W., Kawamura, Y.: Customer behavior pattern discovering with web mining. In: Proceedings of Asia Pacific web conference, Hangzhou, China, pp. 844–853 (2004)
47. Pabarskaite, Z., Raudys, A.: A process of knowledge discovery from web log data: Systematization and critical review. Journal of Intelligent Information Systems 28, 79–104 (2007)