Analysis of Users' Web Navigation Behavior using GRPA with Variable Length Markov Chains

Bindu Madhuri. Ch[1], Dr. Anand Chandulal.J[2], Ramya. K[3] and Phanidra.M[4]

[1]Sr.Assistant Professor, Department of Computer Science & Engineering, GIT, Gitam University, INDIA
binducheekati@gmail.com
[2] Professor, Department of Computer Science & Engineering, GIT, Gitam University, INDIA
chandulal@gitam.edu
[3] Student, Department of Computer Science & Engineering, GIT, Gitam University, INDIA
kavuriramya@gmail.com
[4] Student, Department of Computer Science & Engineering, GIT, Gitam University, INDIA
phani.cse29@gmail.com

Abstract

With the never-ending growth of Web services and Web-based information systems, the volumes of click stream and user data collected by Web-based organizations in their daily operations has reached enormous proportions. Analyzing such huge data can help to evaluate the effectiveness of promotional campaigns, optimize the functionality of Web-based applications, and provide more personalized content to visitors. In the previous work, we had proposed a method, Grey Relational Pattern Analysis using Markov chains, which involves to discovering the meaningful patterns and relationships from a large collection of data, often stored in Web and applications server access logs, proxy logs etc. Herein, we propose a novel approach to analyse the navigational behavior of User using GRPA with Variable-Length Markov Chains. A VLMC is a model extension that allows variable length history to be captured. GRPA with Variable-Length Markov Chains, which reflects on sequential information in Web usage data effectively and efficiently, and it can be extended to allow integration with a Web user navigation behavior prediction model for better Web Usage mining Applications.

Keywords

Web usage mining, Grey Relational Analysis, Markov model, Grey System Theory.

1. Introduction

Web usage mining refers to the automatic discovery and analysis of patterns in click stream; it also involves associated data collected or generated as a result of user interactions with Web resources on one or more Web sites. The goal is to capture, model, and analyze the behavioral

patterns and profiles of users interacting with a Web site. The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common needs or interests. Web usage mining is also known as web log mining. It is the process of discovering and interpreting Patterns of users' accessing the web by mining the web log data.

Generally user page visits are sequential in nature. Markov chains have been used to model user sequential navigational behavior on the web site; because the main functionality of markov chain is that the present state always depends upon the previous state. The concept of Grey Relational Pattern Analysis with Markov Chains [2] is used to analyse Users' navigational behavior. The analyzed behavior of the users is used in application areas for Web usage mining Personalization, System Improvement, Site Modification, Business Intelligence, and Usage Characterization.

The sequential navigation behavior of the user in a Web site represented as a Markov chain, Markov chains and Hidden Markov Models have been enormously successful in sequence matching/generation. Markov chains have many attractive properties. They can be easily estimated statistically. Since the Markov chain model is also generative, navigation tours can be automatically derived. The Markov chain model can also be adapted on-the-fly with additional user navigation information. When used in conjunction with a web server, the same model can be used to predict the probability of seeing a link in the future given a history of accessed links. A discrete Markov chain model can be defined by the tuple $\langle S, A, \lambda \rangle$ **S** is corresponds to the state space; **A** is a matrix representing transition probabilities from one state to another. λ is the initial probability distribution of the states in **S.** The fundamental property of Markov model is the dependency on the previous state. If the vector **s[t]** denotes the probability vector for all the states at time't', then:

$$\hat{S}(t) = \hat{S}(t-1)A \quad \text{---------- (1)}$$

If there are 'n' states in our Markov chain, then the matrix of transition probabilities **A** is of size n x n. Markov chains can be applied to web link sequence modeling. The Markov Chain model consists of a (sparse) matrix (compressed to an appropriate form) of state transition probabilities, and the initial state probability vector. These are stored in the form of both counts and probabilities.

The grey system theory(GST) initiated by Deng [44] can perform grey relational analysis for sequences and is mainly utilized to study uncertainties in system models, analyze relations between systems, establish models, and further make forecasts and decisions. The GRA is an important method in the grey system theory. Recently, the GRA has become very noted in a number of areas such as manufacturing, transportation, and the building trade. In the grey system theory, Yeh et al [31] proposed that the GRA is essentially believed to have captured the similarity measurements or relations in a system .With a given reference sequence and a given set of comparative sequences, the GRA can be used to determine the grey relational grade (GRG) between the reference and each element in the given set. It is the best comparative that can be found by analyzing the resultant GRGs. In other words, the GRA can be viewed as a measure of similarity for finite sequences. It can also be used as a measure of the absolute point-to-point

distance between sequences. It generalizes [2] the concept of grey relational analysis to develop a technique, called grey relational pattern analysis associated with Markov chains for sequential web data, which is used for analyzing the similarity between given patterns. Based on this technique, a clustering algorithm" Grey Clustering algorithm for Sequential Data" is proposed to finding cluster of a given data set .The problem is to determine the optimal number of clusters . For that purpose they developed an evaluation framework in which the Sum of Squared Error (SSE) is calculated to get the efficiency of proposed algorithm.

In the context of Web applications like Personalization, Business Intelligence, and System Improvement etc web usage mining techniques have been utilized to take advantage of the data collected as a result of users' interactions with the web site [44]. Herein, we focus on the problem of building models to represent past users behavior, which enables us to analyze users' behavior. When users click on a link in a web page, submit a query to a search engine or access a wireless network they leave a trace behind them that is stored in a log file. The information stored in the log file for each user click will include items such as a time-stamp, identification of the user (for example, an IP address, a cookie or a tag, IP address+user agent), the user's location, query terms entered and further click stream data, where appropriate. (We use the term 'click' generically to mean a click on a link, a query submission or an access to a network.) Thus the log file contains an entry for each click and can be preprocessed into time-ordered sessions of sequential clicks. In [26] the authors present a study to evaluate heuristics to reconstruct sessions from server log data, known as sessionising. They show that sessions can be accurately inferred for web sites with embedded session identification mechanisms, that time-based reconstruction heuristics are acceptable when cookie identifiers are available, and that referrer-based heuristics should be used when cookie identifiers are not available. Higher-order Markov models have been widely used for modeling user records. For the task, we have proposed a Variable Length Markov Chain (VLMC) method [12] [1], which is an extension of a Markov chain that allows variable length history to be captured [13]. We note that, we have previously proposed (i) a novel approach to analyze the navigation behavior of Users' using GRPA [2] associated with Markov Chains (ii) a method to find the accuracy with which the model represents a collection of sessions.

This paper is organized as follows. Section 2 provides related literature .The next section provides methodology to analyze the Behavior of User using GRPA with Variable Length Markov Chains. Experimental results are provided in Section 4 followed by conclusions in Section 5.

2. Background

Cluster analysis on Web usage data is widely used in various applications, as shown in Table 1.We provide a focused review of these streams of literature. For the task of clustering Web users or user sessions, many studies have adopted a vector model with Boolean-based (visit/no visit) or frequency-based (the number of times a page is visited) representations of Web usage [11] [37] [8] [25] [43]. For example, Yan et al. [43] used the leader algorithm with frequency-based representation of Web usage. Each user session is represented by n-dimensional feature vector and the degree of interest of the users in a specific Web page is calculated based on the number of times the page has been accessed and the amount of time the user spent on the page. Then, based on the degree of interest, clusters of similar sessions are formed. Although Boolean-based or

frequency-based clustering methods are relatively easy to apply and are widely applied, and accepted these methods have drawbacks since information about multiple visits to the same page may be lost and users' transitions from one page to another cannot be reflected. The effects of different representation schemes – usage-based (UB), frequency-based (FB), viewing-time based (VTB) and visiting-order based (VOB) – were examined by [34]. They applied a matrix-based clustering algorithm to cluster user sessions and discovered that, on an average, the number of clusters using VOB scheme is always greater than that using other representation schemes, whereas the number of clusters formed with UB is always the smallest among the four schemes.

The method of using a sequence of visits in order to cluster Web users is often seen in several Web usage mining studies. Shahabi et al., [41] conducted an experiment to evaluate the performance of their path clustering method that fully considers Web users' navigation path and viewing time. Shahabi and Banaei Kashani [24], introduced a model with feature-matrices (FM) for use in discovering and interpreting users' access patterns. The FM model, which is a generalization of the vector model, allows the capture of various spatial and temporal features of Web usage data and enables the use of vector-based distance measures. Kim [4], who used self-organizing map (SOM) for clustering Web user sessions with sequence-dependent representation, asserted that sequence-dependent representation can overcome the limited representational power of Boolean or frequency-based representation. [30] [19] applied Ward hierarchical clustering method with non-vector-based distance measure which is called the sequence alignment method (SAM) and a multidimensional sequence alignment method (MDSAM) before their navigation pattern extraction step. Experiments on a real data set indicate that SAM better represents behavioral characteristics of Web users regarding content and order of Web page visits than the traditional Euclidean-distance measure. However, the pair-wise similarity comparison required for hierarchical clustering methods brings a scalability issue, and thus these experiments were carried out with a relatively small dataset. Kumar et al., [5] proposed using a non-vector-based similarity measure called S^3M for clustering Web user sessions. S^3M Considers both composition of pages visited and the order of visits. [5] Showed in their experiments using two real data sets that the SeqPAM algorithm, a variant of partition around medoids (PAM) that utilizesS^3M, performs better than PAM using frequency-based cosine similarity measure due to the consideration of sequential information. [5] also proposed a hierarchical clustering algorithm that utilizes S^3M. They proposed using a rough set theory-based algorithm developed by De and Krishna [18] along with S^3M similarity. They argue that rough clustering, which allows fuzzy membership (i.e., a session can belong to more than one cluster), can provide better interpretations of different navigation patterns of Web users, unlike traditional clustering algorithms that require a clear fit into one group.

Chen et al.,[40] derived an algorithm to extract maximal forward references (page visits) from a Web server log under the assumption that backward reference by Web users is only made due to the ease of navigation, not due to their navigation behavior. Yang and Parthasarathy [28] used temporally constrained association rule mining to capture Web users' access patterns and used them for future access predictions. Their approach was based on the premise that recent page access records have great influence on future page accesses. In other general sequence mining studies, sequential pattern mining techniques are also used to reduce the computational complexity and to produce meaningful clusters [33] [9]. In order to address the computational complexity arising from pair-wise similarity comparison, Guralnik and Karypis [33] presented a sequence mining method which does not require a similarity matrix for clustering. The idea is to

capture the sequential nature of various data sequences and represent each data sequence into a new feature space so that a computationally efficient vector-based algorithm such as K-means algorithm can be used. They found that the feature-based approach achieved reasonably good clustering results when compared to similarity based approaches, all this while significantly reducing computational complexity. Another approach dealing with complexity, called ApproxMAP, was introduced by Kum et al., [21]. Conventional sequence mining methods which intend to find exact patterns in a complete set of sequence often suffer from generating a voluminous number of short trivial patterns. ApproxMAP identifies patterns approximately shared by many sequences called consensus patterns so that it substantially reduces the number of trivial patterns while providing more accurate and informative insights into sequential data. Regardless of representation scheme, many clustering algorithms are based on a distance measure; the distance is vector-based such as the Euclidean or non-vector-based such as SAM. In contrast to these distance based clustering methods, Cadez et al., [35] used model-based clustering algorithms for the visualization of Web usages. They applied an Expectation-Maximization (EM) algorithm in order to cluster user sessions while learning a mixture of Markov models and argued that EM is easy to implement and memory-efficient for clustering Web users. Smyth, [42] also developed a general sequence clustering approach, using hidden Markov models (HMMs). Another non-distance-based clustering method called matrix clustering was applied by Oyanagi et al., [22] to find sequential patterns in Web server logs. A new algorithm called the ‘‘Ping-Pong” algorithm was used first to extract dense sub matrices from a binary matrix, in which each element indicates the occurrence of visits from one page to another. Then, by synthesizing the resulting sub matrices (clusters), a super sequence is extracted that represents sequential patterns graphically.

Given the fact that a wide variety of clustering algorithms, representation schemes, and distance measures are available for the purpose of clustering Web users, performance comparisons of these choices also appear in a few Web usage mining studies. As mentioned above, Xiao et al.,[34] compared different representation schemes, and Kumar et al.,[5] [6] and Hay et al.,[30] compared performances of their proposed method with another method or distance. Shahabi et al., [41] conducted an experiment to evaluate the performance of their path mining method that fully considers Web users’ navigation path and viewing time. Although the number of individual users (Web user clusters) and the number of paths (user sessions) were limited in their experiment to 10 and 940, respectively, the K-means clustering algorithm in combination with their sequence-aware cosine distance measure was able to recover the original clusters 73–90% in terms of correct probabilistic prediction of the next path. Shahabi and Banaei-Kashani [24] compared the performances of using Euclidean, projected Euclidean, and cosine distance measures using precision and recall metrics and showed that projected Euclidean is the best choice in their dynamic clustering problem context. Martin-Guerrero et al., [8] compared the performances of various clustering algorithms (K-means, fuzzy c-means, hierarchical clustering, expectation-maximization, and SOM). Their simulation result shows that for a very simple Web site, using the K-means algorithm was adequate. But, for more complex Web sites, SOM was the most suitable algorithm. Kuo et al., [11] compared the performance of K-means, ART2, and various hybrid algorithms that are combinations of ART2, SOM, and genetic K-means. The number of misclassifications among proposed clustering methods was calculated to measure the robustness of the algorithms. Their results suggest that using hybrid algorithms result in better performance than using a single clustering method such as K-means alone or ART2 alone. These comparative studies give insights useful for choosing one method over another. They, however, were often

carried out with a limited set of Web site scenarios or in less realistic settings; therefore, in other Web site scenarios, the conclusions of the comparative studies may not be valid. While Markov models are highly regarded as Web usage mining methods M. Deshpande et al., [14] , J. Zhu et al., [32], J. Pitkow et al [39], R.R. Sarukkai, [38] ,R. Sen et al., [23] ; I. Cadez et al., [35] combining Markov models with sequence-based clustering has rarely been attempted, with the exception of Yang et al.,[28] in which Markov chain's transition matrix-based representation and K-means algorithm are used to cluster user sessions for better Web caching and pre-fetching. As shown in Table 1, while sequence-based clustering is consistently highly regarded for use in Web usage mining, a systematic evaluation of these methods was rarely addressed.

In most comparisons done in clustering studies, either the number of clustering methods being compared is limited or the size of the problem, such as the number of sessions, is small. This is partly due to dimensional complexity resulting from sequential data representation or scalability issues arising mainly from existing sequence-based clustering methods. Based on our review of current literature, we were led to conclude that developing a general, sequence-based clustering methodology as well as developing an evaluation framework for the systematic comparison of the various techniques developed so far may help to close some obvious gaps. Sungjune Park et al., [3] proposed a new experimental framework and ANN-enhanced K-means algorithm based on Markov models, which consider sequential information in Web usage data effectively and efficiently, and it can further be extended to allow integration with a Web user navigation behavior prediction model for better Web personalization and System improvement. Bindu et.al, [2] proposed a novel approach, demonstrated in their experiments using three real data sets that the Unsupervised Clustering algorithm, that utilizesS^3M performs better than K-Means algorithm using Transition-based S^3M similarity measure due to the consideration of sequential information as a Markov chains and they used the concept of Grey System Theory (GST), which determines the relationship between the sequential patterns using Grey Relational Pattern Analysis (GRPA). Our methodology, based on GRPA with Variable Length Markov models, considers sequential information in Web usage data effectively and efficiently, and it can be extended to allow integration with a Web user navigation behavior prediction model for better Web Usage mining Applications.

3. Methodology

3.1 VLMC (Variable Length Markov Chain)

A user navigation session within a Web site can be represented by the sequence of pages requested by the user. First-order Markov models have been widely used to model a collection of user sessions. In such context, each Web page in the site corresponds to a state in the model, and each pair of pages viewed in sequence corresponds to a state transition in the model. A transition probability is estimated by the ratio of the number of times the transition was traversed to the number of times the first state in the pair was visited. Usually, artificial states are appended to every navigation session to denote the start and finish of the session.

A first-order Markov model is a compact way of representing a collection of sessions, but in most cases, its accuracy is low S. Jespersen et al., [20] which is why extensions to higher order models are necessary. In a (nonvariable) higher order Markov model, a state corresponds to a fixed sequence of pages [36] and a transition between states represents a higher order conditional probability. For example, in a second order model, each state corresponds to a sequence of two page views. The serious drawback of fixed higher order Markov models is their exponentially large state space compared to lower order models. A VLMC is a model extension that allows variable length history to be captured G. Bejerano [13]. J.Borges and Levene,[12] proposed a method that transforms a first-order model into a VLMC so that each transition probability between two states takes into account the path a user followed to reach the first state prior to choosing the out link corresponding to the transition to the second state. The method makes use of state cloning (where states are duplicated to distinguish between different paths leading to the same state) together with the K-Means clustering technique that separates paths revealing differences in their conditional probabilities. Herein, we make use of a VLMC model to summarize the navigation behavior of the users visiting a Web site. We note that the analysis could be refined by making use of user profile data to group users with similar interests and thereby using an individual VLMC model for each group of users. We also note that it is possible to enhance the model in order to take into account Web page content. For example, each state definition can include a vector of keywords representing the contents of the corresponding Web page. As a result, it would be possible to identify high probability trails that are composed of pages that are relevant to a given topic.

3.1.1 First-Order Model Construction

Sessions	Frequency Count
S P1 P2 P3 P4 F	2
S P1 P3 P5 F	3
S P1 P3 P4 F	2
S P2 P3 P4 F	1
S P2 P3 P5 F	4
S P4 P6 F	9
S P5 P6 F	1

Fig 1.a

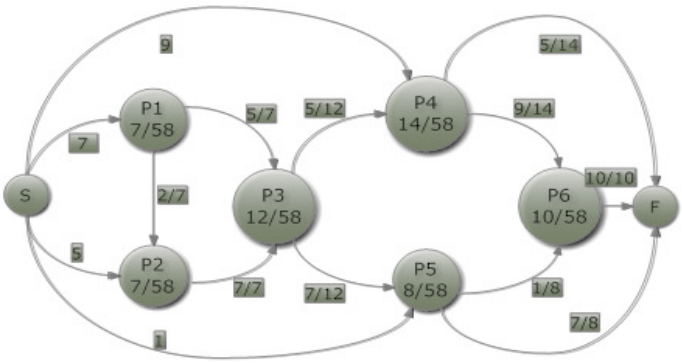

Fig 1.b

Fig. 1.a shows an example of a collection of navigation sessions. We let a session start and finish at an artificial state; Frequency denotes the number of times the corresponding sequence of pages was traversed. Fig. 1.b, presents the first-order model for these sessions. There is a state corresponding to each Web page and a link connecting every two pages viewed in sequence. For each state that corresponds to a Web page, we give the page identifier and the number of times the page was viewed divided by the total number of page views. This ratio is a probability

estimate for a user choosing the corresponding page from the set of all pages in the site. For example, page 4 has 14 page views from a total of 58 page views. For each link, we indicate the proportion of times it was followed after viewing the anchor page. For example, page 5 was viewed 8 times, one of which was at the beginning of a navigation session (the weight of the link from the artificial state S indicates the number of sessions that started in that page). After viewing page 5, the user moved to page 6 in 1 of the 8 times and terminated the session seven times. The probability estimate of a trail is given by the product of the probability of the first state in the trail (that is, the initial probability) and the probabilities of the traversed links (that is, the transition probabilities).

3.1.2 Higher Order Model Construction

The first-order model does not accurately represent all second-order conditional probabilities. For example, according to the input data, the sequence (1,3) was followed five times, that is #(1,3)=5, and sequence (1,3,5) was followed thrice, that is, #(1,3,5)=3. Therefore, the probability estimate for viewing page 5 after viewing 1 and 3 in sequence is p(5|1,3)= #(1,3,5)/ #(1,3)=3/5. The error of a first-order model in representing second-order probabilities can be measured by the absolute difference between the corresponding first- and second-order probabilities. For example, for state 3, we have that | p (5|1, 3) – p (5|3) | = |3/5– 7/12 |=0.017; thus, state 3 is not accurately representing second-order conditional probabilities. The accuracy of transition probabilities from a state can be enhanced by separating the in paths to it that correspond to different conditional probabilities. We fine-ttune the accuracy in the example by cloning state 3 (that is, creating a duplicate state 3') and redirecting the link (2, 3) to state 3'. The weights of the out links from states 3 and 3' are updated according to the number of times the sequence of three states was followed .For example, since # (1,3,4) = 2 and #(1,3,5) = 3 in the second-order model, the weight of the link #(3,4) is 2 and the weight of #(3,5) is 3. (Note that, according to the input data, no session terminates at page 3 when the user has navigated to it from page 1.) The same method is applied to update the out links from the clone state 3'.

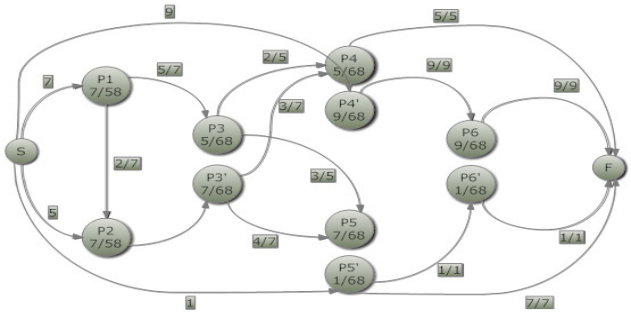

Fig 2

Fig. 2 shows the resulting second-order model after cloning four states in order to accurately represent all second-order conditional probabilities. In the extended model given in Fig. 2, all the

out links represent accurate second-order probability estimates. The probability estimate of the trail # (1, 3, 5) is now 7/58.5/7.3/5 = 0.051. The probability estimate for trail (3, 4) is 5/68.2/5 + 7/68.3/7 = 0.0735, which is equal to the first-order estimate. Therefore, the second order model accurately models the conditional second order probability estimates while keeping the correct first-order probability estimates. In order to provide control over the number of additional states created by the method, we make use of a parameter γ that sets the highest admissible difference between a first order and the corresponding second order probability estimate. In a first-order model, a state is cloned if there is a second-order probability whose difference from the corresponding first-order probability is greater than γ. Alternatively, we interpret γ as a threshold for the average difference between the first-order and the corresponding second-order probabilities for a given state. In the latter, the state is cloned if the average difference between the first and second-order conditional probabilities surpasses γ. When γ is measuring the maximum probability of divergence, we will denote it by γ_m, and when it is measuring the average probability divergence, we will denote it by γ_a.

3.2 Grey Relational Pattern Analysis associated with Variable Length Markov Chains for navigational data

Understanding the behavior of Web site visitors navigating through a site is an important step in the process of improving the quality of service of that site. A model of past user navigation behavior can be used to identify frequent usage patterns that can provide insights on how to improve the Web site design and structure in order to satisfy the visitors needs. In addition, being able to predict the near future navigation intentions of an individual user will enable the provision of pages adapted to the recent behavior of this user and the use of past behavior of other visitors to guide the user in the search to satisfy his information needs. Markov models are well suited for modeling user Web navigation data because they are compact, simple to understand and motivate, expressive, and based on a well-established theory. Commercial tools for log data analysis usually discard the information concerning the order in which page views occurred in a session, and the same can be said about the techniques using association rules methods. On the other hand, variable-length Markov chain (VLMC) models provide the probability of the next link chosen when viewing a Web page while taking into account the trail followed to reach that page.

The Grey Relational Pattern Analysis associated with Markov Chains for navigational data, Bindu et al.,[2] can be considered as it measures similarity between elements with the same index of patterns. This concept can be used for the analysis of the pattern relation. In order to analyze the pattern relation in a similar manner, GRPA with Variable Length Markov Chains is used to analyze the users' behavior. Our methodology, based on GRPA with Variable Length Markov models, considers sequential information in Web usage data effectively and efficiently, and it can be extended to allow integration with a Web user navigation behavior prediction model for better Web Usage mining Applications.

4. Experimental results

The purpose of this Section is to show the performance and effectiveness of our method on various data sets. All the experiments (cti, msnbc and msweb datasets) were carried out on a 2.4 GHz, 256 MB RAM, Pentium-IV machine running on Microsoft Windows 7 and the code was developed on .NET Framework 4.0.

4.1 Web log preprocessing results

In the preprocessing procedure, the original web server log data are cleaned, formatted, and finally grouped into meaningful user sessions. The following table 1 presents some statistics of the experimental data set, which includes both training and testing sets obtained after the preprocessing operations. We can see that for the training dataset, 1500 clean entries are extracted having 500 different users who accessed the web server in two months duration. In this period, 243 web pages were visited, and 134 of them were accessed at least 10 times. Later on user sessions are identified by using the time oriented heuristic method, based on a 30-minute threshold of the session duration. On a whole, totally 312 sessions were identified from the training set by the session duration based method.

Table 1. Dataset Description.

Attributes	Training set	Testing set
Total access entries	22600	18000
Clean access entries	1500	1000
Different access users	500	356
Accessed web pages(total)	243	198
Accessed web pages(≥10 times)	134	106
Identified sessions	312	231

Herein, our goal is to analyze the behavior of the users to provide web Personalization, System Improvement, Site Modification, Business Intelligence, and Usage Characterization. User navigational paths are identified based on the user requests and the referrer pages in order to analyze the behavior of the users, generally these navigations are in sequential manner. After session identification we further preprocessed our dataset where the root pages were considered in the page view of a session. This preprocessing step resulted in total of 8 categories namely, news, admissions, advising, courses, people, research, resources, shared. These page views were given numeric labels as 1 for news, 2 for admissions and so on. Table 2 shows the complete list of numeric coded Web pages. Figure 2 shows the sample Web navigation data. Each row describes the hits of a single user. Comparing very long sessions with small sessions would not be meaningful, hence we considered only sessions of length between 3 and 5. Finally, we took 312 user sessions for our evaluation as shown in Table 2:

Table 2: Example data set

Webpage name	Number coding
News	1
Admissions	2
advising	3
courses	4
People	5
research	6
resources	7
shared	8

.

The concept of Markov models is used for modeling user Web navigation data, as they are compact, simple to understand, expressive, and based on a well-established theory. The page views occurred in a session are represented as a Markov model (first, higher order and VLMC). Herein, we make use of a VLMC model to summarize the navigation behavior of the users visiting a Web site. We note that the analysis could be refined by making use of user profile data to group users with similar interests and using an individual VLMC model for each group of users. We also note that it is possible to enhance the model in order to take into account Web page content. For example, each state definition can include a vector of keywords representing the contents of the corresponding Web page. As a result, it would be possible to identify high probability trails that are composed of pages that are relevant to a given topic. The VLMC navigation sequences are the input to the Grey Clustering algorithm. We compare the result obtained using the grey clustering algorithm with that obtained using the k-means method, in which c is the number of clusters to be predetermined. The initial threshold starts from 0.1 and increases by 0.01 for an optimal solution. In the first learning iteration of the grey clustering algorithm, the threshold value has to choose, which is equal to the initial threshold. Then, the threshold increases also by 0.01 for the next learning iteration for a clustering result.

4.2 Experimental Results on pilot Dataset

200 Web transactions are chosen arbitrarily from the msnbc dataset and pilot experiments are performed on them. The sum of squared error measure is used to find the optimal number of clusters in the Grey clustering technique. Different representation schemes have been employed (frequency-based, transition-based, and precedence based), and tested using the K-means algorithm with Euclidean distance measure (KME). Performance of transition-based representation scheme is found to be optimized in all cases where as frequency- based representation had the worst results in terms of mean SSE measure. It can be explained that, the Markov model we assumed when generating sequences can be best represented by the transition-based sequence matrix, and thus transition-based representation performs better than precedence-

based representation. However, if we had assumed a different model for Web users' behaviors which emphasized precedence, the result would have been different.

4.3 Experimental Results on msnbc Dataset

4.3.1 Dataset Description:

We conducted some pilot experiments on the real datasets. The first data set from the UCI dataset repository (http://kdd.ics.uci.edu/) that consists of Internet Information Server (IIS) logs for *msnbc.com* and news-related portions of *msn.com* for the entire day of September 28, 1999 (Pacific Standard Time). Each sequence in the dataset corresponds to page views of a user during that twenty-four hour period. Each event in the sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail but they are recorded at the level of page categories as determined by the site administrator. There are 17 page categories, namely, "front page," " news," "tech," "local," "opinion,""on-air," "misc," "weather," "health,""living," "business," "sports," "summary,""bbs" (bulletin board service), "travel,""msn-news," and "msn-sports." Each page category is represented by integer label. For example, as shown in figure 3, "frontpage"is coded as 1, "news" as 2, "tech"as 3, etc. Each row describes the hits of a single user. For example, the fourth user hits "frontpage" twice, and the second user hits "news" once and so on .In the total dataset, the length of user sessions ranges from 1 to 500 and the average length of session is 5.7.

Fig 3: Msnbc Data set

Example msnbc web navigation data

T1: on-air misc misc misc on-air misc
T2: news sorts tech local sports sports
T3: bbs bbs bbs bbs bbs bbs
T4: frontpage frontpage sports news news local
T5: on-air weather weather weather sports
T6: on-air on-air on-air on-air tech bbs
T7: frontpage bbs bbs frontpage frontpage news
T8: frontpage frontpage frontpage frontpage frontpage bbs
T9: news news travel opinion opinion m sn-news
T10: frontpage business frontpage news news bbs

Description of the msnbc dataset

Total Dataset	
Number of users	989,818
Minimum session length	1
Maximum session length	500
Average number of visits per user	5.7

Preprocessed msnbc dataset with 44,062 user sessions was given as input, which is post processed as Markov chains (first order, higher order and the variable length) to the Grey Clustering algorithm. The grey clustering algorithm, which is an unsupervised clustering algorithm, uses the concept of grey relational pattern grades to find the relations between the msnbc sequence patterns. The results are summarized in Table 3. The experimental results are shown in Table 4.

Table 3: Summarized Msnbc dataset results

Threshold Value	Number of Clusters	Minimum Length	Number of Sessions	KME	Grey Clustering
0.75	7	2	2000	0.377	**0.456**
		5	5000	0.456	**0.551**
		7	7000	0.569	**0.591**
	9	2	2000	0.367	**0.423**
		5	5000	0.441	0.361
		7	7000	0.521	0.521
	12	2	2000	0.345	**0.452**
		5	5000	0.521	**0.862**
		7	7000	0.792	**0.945**
0.83	7	2	2000	0.486	0.446
		5	5000	0.609	**0.621**
		7	7000	0.582	0.425
	9	2	2000	0.329	**0.427**
		5	5000	0.426	0.321
		7	7000	0.789	**0.920**
	12	2	2000	0.569	**0.679**
		5	5000	0.789	0.673
		7	7000	0.853	**0.893**

Table 4: Experimental results of Msnbc dataset

Threshold value	No of clusters formed	C1		C2		C3		C4		C5		C6		C7		C8		C9		C10	
		C	SSE	C	SSE	C	SSE	C	SSE	C	SSE	C	SSE	C	SSE	C	SSE	C	SSE	C	SSE
0.915	10	0.926	0.67	0.921	0.157	0.914	0.04	0.918	0.01	0.919	0.001	0.912	0.003	0.915	0.009	0.921	0.018	0.92	0	0.924	0.01
0.92	9	0.926	0.67	0.921	0.157	0.914	0.04	0.918	0.01	0.917	0.045	0.915	0.09	0.919	0.024	0.918	0.018	0.923	0.024	X	X
0.925	6	0.921	0.58	0.923	0.306	0.9178	0.308	0.9173	0.04	0.911	0	0.914	0.01	X	X	X	X	X	X	X	X
0.93	5	0.926	0.76	0.916	0.11	0.917	0.04	0.912	0.06	0.914	0.01	X	X	X	X	X	X	X	X	X	X
0.935	5	0.921	0.62	0.916	0.18	0.914	0.038	0.917	0.04	0.918	0.01	X	X	X	X	X	X	X	X	X	X
0.94	4	0.926	0.76	0.916	0.11	0.917	0.048	0.915	0.02	X	X	X	X	X	X	X	X	X	X	X	X
0.945	3	0.926	0.76	0.914	0.05	0.915	0.01	X	X	X	X	X	X	X	X	X	X	X	X	X	X
0.95	2	0.926	0.76	0.914	0.05	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

4.4 Experimental Results on CTI Dataset

4.4.1 Dataset Description:

The second data set, *cti*, is from a university Web site log and was made available by the authors of Mobasher (2004) and Zhang et al. (2005). The data is based on a random collection of users visiting university site for a 2-week period during the month of April 2002. After data preprocessing, the filtered data consisted of 13745 sessions and 683 pages. We further preprocessed *cti* dataset where the root pages were considered in the page view of a session. This preprocessing step resulted in total of 16 categories namely, search, programs, news, admissions, advising, courses, people, research, resources, authenticate, cti, pdf, calendar, shared, forums, and hyperlink. These page views were given numeric labels as 1 for search, 2 for programs and so on. Table 3 shows the complete list of numeric coded Web pages. Figure 4 shows the sample cti Web navigation data. Each row describes the hits of a single user. For example, the seventh user hits "research" twice then "course" followed by "news" twice. The session length in the dataset ranges from 2 to 68.Since comparing very long sessions with small sessions would not be meaningful, we considered only sessions of length between 3 and 7. Finally, we took 5915 user sessions for our Experimentation.

Fig 4: Cti Data set

Example cti web navigation data

news people
programs programs admissions programs courses
resources forums
courses courses courses courses
courses people
hyperlink news
reasearch research courses news news
authenticate cti programs cti cti
authenticate cti news
people admissions cti cti admissions admissions people people

Number coding of Web pages for cti dataset

Web Page Name	Number Coding	Web Page Name	Number Coding
Search	1	Resources	9
Programs	2	Authenticate	10
News	3	cti	11
Admissions	4	Pdf	12
Advising	5	Calendar	13
Courses	6	Shared	14
People	7	Forums	15
Research	8	Hyperlink	16

Preprocessed cti dataset with 13745 user sessions was given as input, represented as Markov chains (first order, higher order and the variable length) to the Grey Clustering algorithm. The Grey clustering algorithm, which is an unsupervised clustering algorithm, uses the concept of grey relational pattern grades to find the relations between the msnbc sequence patterns. The results are summarized in Table 5.

Table 5: summarised Cti dataset results

Threshold Value	Number of Clusters	Minimum Length	Number of Sessions	KME	Grey Clustering
0.75	2	2	1000	0.397	**0.412**
		5	2000	0.486	**0.499**
		7	4000	0.598	0.598
	4	2	1000	0.307	**0.398**
		5	2000	0.412	0.401
		7	4000	0.545	**0.565**
	7	2	1000	0.285	**0.321**
		5	2000	0.398	**0.412**
		7	4000	0.692	**0.705**
0.98	2	2	1000	0.477	**0.496**
		5	2000	0.519	**0.621**
		7	4000	0.582	0.560
	4	2	1000	0.429	**0.444**
		5	2000	0.526	0.527
		7	4000	0.589	**0.604**
	7	2	1000	0.469	**0.470**
		5	2000	0.689	**0.701**
		7	4000	0.753	0.693

4.5 Experimental Results on msweb Dataset

4.5.1 Dataset Description:

The third dataset (MSWEB) was obtained from the UCI KDD archive http://kdd.ics.uci.edu/databases/msweb/msweb.html) and records the areas within www.microsoft.com that users visited in one-week time frame during February 1998. Two separate data sets are provided, a training set and a test set. Table 2 summarizes the characteristics of the data set. After data preprocessing, the filtered data consisted of 32711 sessions and 285 pages. We further preprocessed MSWEB dataset where the root pages were considered in the page view of a session. This preprocessing step resulted in total of 20 categories namely “library”, “developer”, “home”, “finance”, “repository”, “gallery”, “catalog”, “mail”, “ads”, “education”, “magazine”, “support”, “ms”, “technology”, “search”, “country”, “business”, “entertainment”, “news”, “feedback”.

Fig 5: Msweb Data set

Training Dataset	
Total number of sessions	32711
Total number of pages	285
Total number of requests	98654
Average Session Length	3.0

Test Dataset	
Total number of pages	236
Total number of requests	15191
Total number of sessions	5000
Average Session Length	3.0

Above indicated are: the number of pages occurring in the log file, the total number of requests recorded, and the total number of sessions derived from each data set. In MSWEB data sets, the contents of a test set is representative of the corresponding training set that the average session length in the training sets is very close to the corresponding values in the test sets. We note that not all pages in a given training set are in the corresponding test set, meaning that test sets do not cover all the web pages that are present in the training sets.

Preprocessed msnbc dataset with 32711 user sessions was given as input, which is post processed as Markov chains (first order, higher order and the variable length) to the Grey Clustering algorithm. The grey clustering algorithm, which is an unsupervised clustering algorithm, uses the concept of grey relational pattern grades to find the relations between the msnbc sequence patterns. The results are summarized in Table 5.

Table 5: summarised Msweb dataset results

Threshold Value	Number of Clusters	Minimum Length	Number of Sessions	KME	Grey Clustering
0.85	5	3	3000	0.411	**0.578**
		6	7000	0.476	**0.512**
		9	8000	0.535	0.509
	7	3	3000	0.397	0.380
		6	7000	0.451	**0.482**
		9	8000	0.498	**0.499**
	9	3	3000	0.385	**0.402**
		6	7000	0.421	**0.493**
		9	8000	0.492	0.478
0.92	5	3	3000	0.456	**0.496**
		6	7000	0.523	**0.528**
		9	8000	0.654	0.615
	7	3	3000	0.419	**0.478**
		6	7000	0.526	**0.621**
		9	8000	0.689	**0.720**
	9	3	3000	0.369	**0.379**
		6	7000	0.589	0.573
		9	8000	0.653	**0.693**

5. Conclusion

In this study, a general sequence-based Web usage mining methodology was developed by utilizing new sequence representation schemes in association with Variable Length Markov models. While the previous work in sequence mining focused more on sequential pattern discovery, our study addresses the use of sequential information in clustering Web users. Our new methodology was tested by examining cluster Sum of Squared Error (SSE) as a clustering performance measure on different real data sets. It was confirmed that identifying Web user clusters through sequence-based clustering methods helps recover the Web user clusters correctly. While demonstrating the use of our experimental evaluation framework to compare sequence-based clustering methods, we used distance

Similarity measure named S^3M.The Unsupervised clustering algorithm with Variable length Markov chains using S^3M Similarity measure, is employed which utilize a concept of Grey System Theory, in which the Grey relational Analysis determines the relationship between the sequences. And the proposed algorithm preformed better than other clustering methods. Understanding the behavior of Web site visitors navigating through a site is an important step in the process of improving the quality of service of that site. A model of past user navigation behavior can be used to identify frequent usage patterns that can provide insights on how to improve the Web site design and structure in order to satisfy the visitors needs. In addition, being able to predict the near future navigation intentions of an individual user will enable the provision of pages adapted to the recent behavior of this user and the use of past behavior of other visitors to guide the user in the search to satisfy his information needs. Moreover, predicting the user's next choice enables the construction of dynamic pages in advance or the provision of a speculative prefetching service that sends, in addition to the requested document, a number of other documents that are expected to be requested in the near future. In our opinion, Markov models are well suited for modeling user Web navigation data because they are compact, simple to understand and motivate, expressive, and based on a well-established theory. Commercial tools for log data analysis usually discard the information concerning the order in which page views occurred in a session, and the same can be said about the techniques using association rules methods. On the other hand, variable-length Markov chain (VLMC) models provide the probability of the next link chosen when viewing a Web page while taking into account the trail followed to reach that page.

References

[1] J. Borges and Levene, M. (2010). A comparison of scoring metrics for predicting the next navigation step. November 3, 2010 15:7 WSPC/ws-ijitdm

[2] Ch. Bindu Madhuri and Dr. Anand Chandulal.J, "Analysis of the Navigation Behavior of the Users' using Grey Relational Pattern Analysis with Markov Chains". International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5402-5412

[3] Sungjune Park, Nallan C. Suresh, Bong-Keun Jeong "Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm", Data & Knowledge Engineering 65 (2008) 512–543

[4] Y. Kim, Weighted order-dependent clustering and visualization of Web navigation patterns, Decision Support Systems 43 (4) (2007) 1630–1645.

[5] P. Kumar, R.S. Bapi, P.R. Krishna, SeqPAM: a sequence clustering algorithm for Web personalization, International Journal of Data Warehousing and Mining 3 (1) (2007) 29–53.

[6] P. Kumar, P.R. Krishna, R.S. Bapi, S.K. De, Rough clustering of sequential data, Data and Knowledge Engineering 63 (2) (2007) 183–199.

[7] B. Mobasher(2007). The adaptive web - methods and strategies of web personalization. Chapter Data mining for web personalization, pages 90 – 135. Springer.

[8] J.D. Martin-Guerrero, A. Palomares, E. Balaguer-Ballster, E. Soria-Olivas, J. Gomez-Sanchis, A. Soriano-Asensi, Studying the feasibility of a recommender in a citizen Web portal based on user modeling and clustering algorithm, Expert Systems with Applications 30 (2) (2006) 299–312.

[9] J. Wang, Xindong Wu, X., & Zhang, C., (2005). Support vector machines based on K-means clustering for real-time business intelligence systems. International Journal of Business Intelligence and Data Mining, 1(1), 54-64, Inderscience Publishers.

[10] Y. Zhang, Xu, G., & Zhou, X. (2005). A latent usage approach for clustering Web transaction and building user profile. ADMA (pp. 31-42).

[11] R.J. Kuo, J.L. Liao, C. Tu, Integration of ART2 neural network and genetic K-means algorithm for analyzing Web browsing paths in electronic commerce, Decision Support Systems 40 (2) (2005) 355–374.

[12] J. Borges and Levene, M. (2005). Generating dynamic higher-order Markov models in web usage mining. In Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), pages 34–45, Porto, Portugal.

[13] G. Bejerano (2004). Algorithms for variable length Markov chain modeling. Bioinformatics, 20:788–789.

[14] M. Deshpande, G. Karypis, Selective Markov models for predicting web-page accesses, ACM Transactions on Internet Technology 4 (2) (2004) 163–184.

[15] F. Masseglia, D. Tanasa, B. Trousse,Webusage mining: Sequential pattern extraction with a very low support, in: J.X. Yu, X. Lin, H. Lu,Y. Zhang (Eds.), Advanced Web Technologies and Applications, Lecture Notes in Computer Science, vol. 3007, 2004, pp. 513–522.

[16] B. Mobasher (2004). Web usage mining and personalization. In M. P. Singh (Ed.), Practical handbook of Internet computing. CRC Press.

[17] D. Oikonomopoulou, M. Rigou, S. Sirmakessis, A. Tsakalidis, Full-coverage Web prediction based on Web usage mining and site topology, in: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), Beijing, China, 2004, pp.716–719

[18] S.K. De, P.R. Krishna, Clustering Web transactions using rough approximation, Fuzzy Sets and Systems 148 (1) (2004) 131–138.

[19] B. Hay, G. Wets, K. Vanhoof, Segmentation of visiting patterns on Web sites using a sequence alignment method, Journal of Retailing and Consumer Services 10 (3) (2003) 145–153.

[20] S. Jespersen, T. Pedersen, and J. Thorhauge, "Evaluating the Markov Assumption for Web Usage Mining," Proc. Fifth ACM Int'l Workshop Web Information and Data Management, pp. 82-89, 2003.

[21] H.-C.M. Kum, J. Pei, W. Wang, D. Duncan, ApproxMAP: approximate mining of consensus sequential patterns, in: Proceedings of the Third SIAM International Conference on Data Mining (SDM), San Francisco, CA, 2003, pp. 311–315.

[22] S. Oyanagi, K. Kubota, A. Nakase, Mining WWW access sequence by matrix clustering, in: O.R. Zaiane, J. Srivastava, M.Spiliopoulou, B. Masand (Eds.), Mining Web Data for Discovering Usage Patterns and Profiles, Lecture Notes in Computer Science,vol. 2703, Springer, Berlin/Heidelberg, 2003, pp. 119–136.

[23] R. Sen, M.H. Hansen, Predicting Web user's next access based on log data, Journal of Computational and Graphical Statistics 12 (2003) 143–155.

[24] C. Shahabi, F. Banaei-Kashani, Efficient and anonymous web-usage mining for Web personalization, INFORMS Journal on Computing 15 (2) (2003) 123–147.

[25] K.A. Smith, A. Ng, Web page clustering using a self-organizing map of user navigation patterns, Decision Support Systems 35 (2) (2003) 245–256.

[26] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa, "A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis," INFORMS J. Computing, no. 15,pp. 171-190, 2003.

[27] Q. Yang, J.Z. Huang, M. Ng, A data cube model for prediction-based Web prefetching, Journal of Intelligent Information Systems 20(1) (2003) 11–30.

[28] H. Yang, S. Parthasarathy, On the use of constrained associations for Web log mining, in: WEBKDD 2002 – Mining Web Data for Discovering Usage Patterns and Profiles, Lecture Notes in Computer Science, vol. 2703, Springer, Berlin/Heidelberg, 2003, pp. 100–118.

[29] J. Yang, W. Wang, CLUSEQ: efficient and effective sequence clustering, in: Proceedings of the 19th International Conference on Data Engineering, Bangalore, India, 2003, pp. 101–112.

[30] B. Hay, G. Wets, K. Vanhoof, Web usage mining by means of multidimensional sequence alignment methods, in: O.R. Zaiane, J.Srivastava, M. Spiliopoulou, B. Masand (Eds.), WEBKDD 2002 – Mining Web Data for Discovering Usage Patterns and Profiles, Lecture Notes in Computer Science, vol. 2703, Springer, Berlin/Heidelberg, 2003, pp. 50–65.

[31] Yeh, M.-F., Chang, J.-C., and Lu, H.-C.: 'Unsupervised clustering algorithm via grey relational pattern analysis', J. Chinese Grey Syst.Assoc, 2002, 5, (1), pp. 17–22

[32] J. Zhu, J. Hong, J.G. Hughes, Using Markov chains for link prediction in adaptive Web sites, in: D. Bustard, W. Liu, R. Sterritt (Eds.), Software 2002: Computing in an Imperfect World, Lecture Notes in Computer Science, vol. 2311, Springer, Berlin/Heidelberg, 2002, pp. 60–73.

[33] V. Guralnik, G. Karypis, A scalable algorithm for clustering sequential data, in: Proceedings of the IEEE International Conference on Data Mining, San Jose, CA, 2001, pp. 179–186.

[34] J. Xiao, Y. Zhang, X. Jia, T. Li, Measuring similarity of interests for clustering web-users, in: Proceedings of the 12th Australasian Conference on Database Technologies, 2001, pp. 107–114.

[35] I. Cadez, D. Heckerman, C. Meek, P. Smyth, S. White, Visualization of navigation patterns on a Web site using model-based Clustering, in: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001), Boston, MA, 2000, pp. 280–284.

[36] J. Borges and Levene, M. (2000). Data mining of user navigation patterns. In Masand, B. and Spiliopoulou, M., editors, Web Usage Analysis and User Profiling, LNAI 1836, pages 92–111.Springer-Verlag, Berlin.

[37] Y. Fu, K. Sandhu, M.-Y. Shih, A generalization-based approach to clustering of Web usage sessions, in: B. Masand, M. Spiliopoulou (Eds.), Web Usage Analysis and User Profiling: International WEBKDD'99 Workshop, San Diego, CA, Lecture Notes in Computer Science, vol. 1836, Springer, Berlin/Heidelberg, 2000, pp. 21–38.

[38] R.R. Sarukkai, Link prediction and path analysis using Markov chains, Computer Networks 33 (1) (2000) 377–386.

[39] J.Pitkow, P. Pirolli, Mining longest repeating subsequences to predict World Wide Web surfing, in: Proceedings of the USITS'99: The 2nd USENIX Symposium on Internet Technologies and Systems, Boulder, Colorado, 1999, pp. 139–150.

[40] M.-S.Chen, J.S. Park, P.S. Yu, Efficient data mining for path traversal patterns, IEEE Transactions on Knowledge and Data Engineering 10 (2) (1998) 209–221.

[41] C.Shahabi, A.M. Zarkesh, J. Adibi, V. Shah, Knowledge discovery from users web-page navigation, in: Proceedings of the Seventh International Workshop on Research Issues in Data Engineering, RIDE'97, April 7–8 1997, Birmingham, UK, 1997, pp. 20–29

[42] P.Smyth, Clustering sequences with hidden Markov models, in: M. Mozer, M. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems, vol. 9, The MIT Press, 1997, pp. 648–654.

[43] T.W.Yan, M. Jacobsen, H. Garcia-Molina, U. Dayal, From user access patterns to dynamic hypertext linking, in: Proceedings of the Fifth International World Wide Web conference, Paris, France, 1996, pp. 1007–1014.

[44] Deng, J.L. Introduction to Grey System Theory, The journal of Grey theory, Vol. 1, Issue 1, 1988, pp. 1-24.