International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

Network Assembling To Identify Network Groupings Using a Fusion Diminution Scheme

Y.Praveen Kumar, A.Swarna, Zaheer Ahmed
VJIT

ABSTRACT— Clustering is the most essential technique to organize similar objects into proper groups based on features in the domain of data mining, machine learning and pattern recognition. In each cluster, objects are more similar to each other on the basis of particular features. Clustering has numerous applications in multiple domains such as information retrieval, data mining, machine learning, pattern recognition, mathematics, medical and bioinformatics. As a result of the unending extension of e-business, there is outstanding contention among relationship to pull in and hold customers. Examinations of the network server logs of these affiliations are essential for getting lots of information into network personalization lead, which can reinforce the arrangement of additionally engaging network structures. Network driven applications are growing step by step and the network has ended up one of the biggest information vaults. Similarity computation calculation among the information objects (network sessions) is mind boggling, however is a critical issue in unsupervised learning. This research is an attempt to overcome these challenges and problems. The objective of this research paper is to introduce a HAC based similarity measure to compute the similarity among the sessions. A HAC based approach is being applied to compute the statistically significant relationship between observed and expected frequencies of the number of pages visited and the time consumed by a user during a session. Also, Hierarchical agglomerative clustering (HAC) technique is proposed to extract useful knowledge from network log. This helps to improve the visualization of network logs and is equally important for networksite designers, developers and owners for the improvements of networksites at each level. Experimental results with two different log files reveal that the proposed similarity measure with HAC algorithm has significantly improved the computation among data objects in network sessions.

Keywords: Hybrid agglomerative clustering, Soft Computing, Network Assembling, Network Usage Mining, Network Log Mining,

I. INTRODUCTION

Network Usage Mining is the disclosure of client get to designs from Network server get to logs [2]. Network Usage Mining investigations aftereffects of client connections with a Network server, including Network logs and database exchanges at a Network networkpage. Network use mining incorporates bunching to discover characteristic gathering of clients or pages, relationship to find the URLs asked for together and investigation of the successive request in which URLs are gotten to. The blend of advancement and the World Wide Network (WWW) has achieved abundance of digitized data and has opened new horizons for the Research gathering to explore electronic data in different estimations. Therefore, the network has turned into a driving data hotspot for the worldwide group. With the progression of time since its commencement in 1990, the network is filling in as a mass travel course for the conveyance of administrations and assets to all parts of the world.

The network is a system of systems of interrelated PCs, while sites and networksite pages give key data to its clients through the network. Sites are propelled on any network server over the network. There are two major issues, which are moving in parallel with the development of the network: (1) there is no understanding of a unified network server over the network and consequently there is no component to catch the client criticism and click history at a unified level, (2) a site can be composed and created with or without resulting standard improvement methodology [3]. This has opened a

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

variety of issues over the network, for example, client conduct investigation, data recovery, prescribed framework, client relationship administration frameworks, profiling, forecast, and hacking.

Network log records are archives that summaries the exercises of clients that have been happened when perusing networkpage. These log records abide in the network server. Network log archives contain information about User name, IP address, Timestamp, Access request, number of bytes traded and User administrator. Examination of these log records gives course lead of the customer. The data set away in the log archives don't present a correct photograph of the customer's gets to the user's click record is the key to investigate user trends and behavior on a specific networksite. The analysis of user click streams is useful in many ways such as networksite management networksite administration, fraud detection, network personalization, information retrieval systems and recommended systems. Due to decentralized network hosting, the user's click record is also decentralized and has no centralized system to capture the user networksite traversing history.

Consequently, we have to rely on network server log to study the user behavior and trends over a networksite. Network Usage Mining (WUM) comprises of three primary strides: preprocessing, information extraction and results examination [4]. The objective of the preprocessing step is to change the crude network log information into an arrangement of client profiles. Each such profile catches a succession or an arrangement of URLs speaking to a client session. Network use information preprocessing abuse an aassortment of calculations and heuristic strategies for different preprocessing undertakings, for example, information cleaning, client recognizable proof, session ID and so forth.

There are three noteworthy sources for network log, for example, proxy network log, customer network log and network server log documents [7]. Every network log source is fragmented and has distinctive upsides and downsides. In the past, majority of the studies reveals that the network server log is utilized and considered as a legitimate source for the investigation of client click streams. In any case it is inadequate as the client may utilize the networksite pages from network cache too [8]. The program store support can break the client succession in network server log record. This issue can likewise be handled by utilizing the network site structure to finish the missing what's more, broken edges. The Network Usage mining (WUM) plays a key part in network site administration and network site organization. It is an expansion of information mining systems to remove the covered up information from network log and has various applications such as feature identification, design revelation, network personalization, recommender systems, frameworks and network user behavior analysis [9].the basis of the similarity measures of all the attributes of data objects of a session. Subsequently, the merged group is again combined with yet another group to form larger clusters on the basis of similarity among the session objects. The merging process is carried until the stopping criterion of the single largest group is obtained.

Network utilization information are unlabeled so they don't contain any class data. The HAC calculations can group comparable client sessions in an effective way in view of the frequencies at which URLs are gotten to amid client sessions. Network client session information typically contain off base, conflicting, and missing data. These shortcomings negatively affect the bunch disclosure process. In this manner, the groups framed won't not be solid and reliable. Be that as it may, the HAC procedures use the fluffy enrollment idea in fluffy sets, which are more vigorous against flaws, so they are more reasonable than customary hard grouping systems for design disclosure in blemished information.

Due to the non-deterministic perusing examples of different network clients, client session information don't have fresh limits and they frequently frame covering bunches [10]. As a result of the covering idea of network client session information, HAC grouping procedures can be connected extremely well to frame covering bunches, where every client session protest can have a place with a few groups with various degrees of enrollment. Additionally, HAC calculations are straightforward, proficient, simple to execute, and they have been utilized broadly to mine network use information. The HAC calculation has the additional favorable position that it is more strong against commotion contrasted and different calculations.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

II. RELATED WORK

Network mining is facing different challenges such as robustness to noise, number of clusters, multi-resolution of the data, mining only good clusters, and efficiency. In their proposed research the hierarchical unsupervised niche clustering algorithm (H-UNC) with robust weights was applied for session clustering. For H-UNC, genetic algorithm (GA) was used to address the robustness issues [14]. The fitness function used for clustering is given in equation 1.

$$f_i = \frac{\sum_{j=1}^{N} w_{ij}}{\sigma_i^2} \quad (1)$$

where Wij is robust scale dispersion measure. The fitness fun ij is the robust weight and $\sigma^2 1$ is the robust scale dispersion measure. The fitness function (equation 1) gives optimum results at the centroid of the cluster. The proposed H-UNC was 2-dimensional and used the Euclidean distance to find the similarity among the sessions. The Euclidean measure is widely criticized due to its nature and its application in network usage mining.

A scalable immune system clustering algorithm for user profiles mining in network log data under single pass was proposed to cluster the logs [15]. The proposed algorithm was inspired by the natural immune system to adopt dynamic changes. The network server was to act like a human body and click streams were marked as antigens. White blood cells (B-cells) detection and destroy system was used to detect the noisy click data in dynamic weighted B-cells (DWB). The weighted influence zone of each profile is calculated in equation 2.

$$\sigma i^2 = \sum^{J} j=1 \ Wij \ dij^2 \ / \ 2$$
$$\sum^{J} j=1 \ Wij \quad (2)$$

The euclidean distance, Cosine and Jaccard measures are not suitable measures for network session clustering due to the nature of user click stream data [16]. He proposed the time based and URL page similarity among the pages visited by different users. For any two network pages visited, the page viewing time was [0, 1] and for matching similarity, the similarity score is 20 and for mismatch and in between the gap, the similarity score is -10. To compute the similarity, dynamic programming was used. The only issue of match and mismatch among the sessions were considered while the similarity must be relative to sessions. Furthermore, hierarchical Assembling was not performed for focused visualization.

$$Stime = \frac{\min(ttimeA, ttimeB)}{\max(ttimeA, ttimeB)} \quad (3)$$

The network user session clustering by applying the agglomerative bunching calculation is proposed to group the network clients. Arrangement score (Sa) and nearby similitude (Sb) are two noteworthy parts to figure the likeness between sessions. The connected dynamic programming on sessions and progressive bunching procedure to pick the outcomes.

The comparability was computed by utilizing the longest normal subsequence (LCS) and connected the grouping calculation to bunch the sessions.

$$Sa(S1, S2) = v / S(m) * M) \quad (4)$$

$$Sim(S1, S2) = Sa * Sb \quad (5)$$

The issue of time spent on a network page is discussed and used for network session clustering. It is very difficult to calculate the proper time utilization on a single page. A page consists of different network objects and each network object has a different worth to different users.

For further details on network objects and network pages, some users may spend more time on that particular page while the other user may not [17]. Consequently, such type of approaches may work for a networksite consisting of a few pages, whereas for the larger networksites this technique is not scalable.

$$S'' = (TLCS^{\alpha} / T^{\alpha} * T^{\beta}LCS / T^{\beta}) ^{1/2} \quad (6)$$

$$S\alpha$$
$$\beta = S' * S'' \quad (7)$$

The significance of similarity measure for network sessions were calculated with the similarity in two steps. In the first step, similarity among the network pages is calculated by tokenizing the pages' URL and by using the longest URL common string.

The string matching criteria stops when the URL of two completely mismatch [18]. For the matching network pages, similarity is marked as 1.0 and for the mismatch pair, it is 0.0. In the second step, the similarity among the network sessions is calculated and matching network pages in two sessions, the

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

matching score is taken as 20 and for mismatch network pages it is -10. The higher the score between the sessions, the higher will be the similarity among the network sessions.

Today, the concept of dynamic network pages is common and their technique is silent. Moreover, the technique is not scalable for larger networksites. Another limitation is that it is not necessary for the network page designer to design the networksite properly and follow the network pages naming conventions.

The clustering technique, whether it is supervised, semi-supervised, or unsupervised, is used to manage the efficiency and accuracy issues [19]. The author categorized the network usage data as heterogeneous because it is composed of different formats such as numerical and categorical. The session time, number of pages visited in a session and data downloaded in a session are numerical, while the pages visited are categorical. To find out the similarity among the network sessions in such a sparse nature of data is a tough task he used a two-step technique to compute the similarity among the sessions. A framework COWES was proposed by the author for the network user clustering based on evolutionary network sessions [20].The similarity among the users is calculated through the fractures and each network user is represented as a set of fractures. User similarity (US) is computed in the range [0,1] in the following equation

$$US(u_1,u_2)=\sum_{k=1}^{n}\delta_k FS_k(u_1,u_2)\Big/\sum_{k=1}^{n}\delta_k \qquad (8)$$

where δk are shared fractures of two users. The clustering was performed by the standard agglomerative algorithm. The two major limitations were discussed for the proposed similarity such as common fractures and the denominator as total shared fractures.

III. FEATURE SELECTION AND SESSION WEIGHT ASSIGNMENT

The user session are mapped as vectors of URL references in the n-dimensional space. The unique URL from the set of preprocessed log file is taken for the experiment. The user sessions taken from the preprocessed log file is taken where each user session is represented as a set of weighted sessions. The weights assigned to the user sessions is represented as a binary and non-binary values which depends on the URL or the some feature of the URL sessions. Let *U = {u1,u2,u3……un}* be the unique URL and the user sessions are represented as *S{s1,s2,……sm}*.The weights of the user sessions are represented as wui

IV. ASSIGNING WEIGHTS TO THE USER SESSION : FUZZY APPROACH

The preprocessing of the session files were done to remove the noise and inconsistent data from the user clusters. The direct removal of the cluster data will result in the loss of significant amount of information when the dataset of the session file is large. To overcome this problem we used the Fuzzy set theoretic approach. Using this method the threshold is specified to remove only the unwanted sessions. The weights are assigned by using the Fuzzy membership function based on the URL accessed by the network users. The session weight using the linear fuzzy membership is carried using the equation as given above.

Algorithm: getLWPs (List SD, double MSLWP)

Input: A set of session-based network data SD; a user-specified minimum support MSLWP.

Output: A set of large network pages for each network user.

Initialize Dataset(D_w)

Initialize weights(W_i)

foreach(D_w in set U={D_{w1}.....D_{wn}})

Initialize the cluster set(C_i)

If($C_i > W_i$)

SD={W_{i+1},D_{w1}}

End

MSLWP=SD+W_i;

End

Assembling issue is a vital stride in WUM process and it is considered as a substantial and solid answer for the accomplishment of WUM. In the event that we have temperamental Assembling, whatever remains of the WUM procedure may create ridicule comes about and at last make the framework blunder inclined what's more, defenseless. The framework may not look for the fancied position in the choice emotionally supportive network. The WUM handle includes various interrelated procedures what's more, these

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

procedures are executed in various stages. A brief portrayal of these WUN steps is as beneath:

A. NETWORK USAGE DATA AND PREPROCESSING

Network utilization information and preprocessing Network server log record is the essential natural information source for WUM strategy and the network get to document is a noteworthy wellspring of crude information. The diverse network server log documents has been talked .Network log is put away in plain content organization (ASCII) and that is a part of the working framework as opposed to a piece of network application. Get to log, operator log, blunder log, and referrer log are usually accessible network sign on network servers. Figure 1 demonstrates a nonspecific preview of the network log that conveys the qualities and significant data about the client crossing click history The log document records the client click stream while the client surfs the site, and because of the utilization of stateless convention HTTP, the log document records all articles (sound, video, pictures, robots) accessible on that solitary page along with the page URL. The greater part of the log sections are unessential for mining method. As the innovation has engaged us to catch a gigantic measure of network information, network log documents are a noteworthy wellspring of network information that store client click streams.

As per log documents contain 60 % immaterial information and that can't be utilized for information mining purposes. In this way we have utilized the college network log documents for our trials. Preprocessing step is essential and of network log record gets to be basic. For precise results, preprocessing is an essential stride in NetworkKDD. The purging was performed to have appropriate information for the WUM procedure. We expelled the sound, video, CSS, robots and crawlers passages because of the outline way of the site. The passages are unessential for the mining reason and must be dispensed with before applying the information mining procedures. These crude passages assume no part in mining and make comes about deride.

The passages, for example, picture records, CSS sheets, scripting, robots, crawlers, sound and video passages are recorded in a network log. Log documents likewise record the authoritative activities for example, overhaul, embed, or erases. Thusly, all these superfluous passages must be evacuated for nature of the WUM prepare.

We hold just valuable and mining required passages. The fruitful passages whose status code = "200" are kept while alternate passages are disposed of. The purifying stride helps us set up the network log for the following steps of WUM process. We applied various cleansing techniques to have a noise free network log file for further processes.

B. NETWORK PERSONALIZATION AND HIERARCHICAL CLUSTERING

Log Assembling is performed on the premise of IP address, nonetheless, clients have the choice to utilize diverse programs, diverse working frameworks and distinctive forms of HTTP. Clients likewise have an alternative to utilize sites from diverse topographical areas. These minor changes can be dealt with as hazard relief for the client examination furthermore, concentrating on the client patterns. In this proposed look into, we customized client's crosses, which are special what's more, unique in relation to the past snap history. This personalization recognizes the business rules for a particular site. For various leveled bunching of network log, we ascertained the quantity of site pages went by the client in a session and subsequent to performing preprocessing, we acquired 1987 sessions. In addition, we ascertained the chi-square values in view of the parameters of a number of networksite pages and a session time in every session. The chi-square estimation of every session is figured with each other session and the most noteworthy chi-square esteem demonstrates the most grounded connection between's these two sessions. Assuming more than one sessions have the same higher esteem, then the in the first place event is viewed as a more suitable combine of related sessions. This is the principal level chain of importance. We moreover registered the normal of the most related sets for the estimation of next chain of importance level and for the tallness of related session in dendogram. We connected the accompanying proposed calculation (Figure 1) for chi based progressive Assembling of network log.

IV. RESULTS AND DISCUSSION

Site log documents contain touchy information and site proprietors for the most part waver to uncover the site log records. Because of this prevention

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

banks, online closeouts and network based shopping site proprietors don't impart their log documents to scientists. For the present study, distinctive site log documents of two unique colleges were chosen. Network log 1 contains a sum of 60302 client click steams in four days and network log 2 contains an aggregate of 65536 client navigates in one day as network log records contain a gigantic measure of insignificant sections because of site structure, before playing out the trial we connected the information preprocessing strategy to set up the information for the genuine analysis. Amid preprocessing ventures around 40 % of passages were expelled as unessential. After the preprocessing stage, Assembling step was performed to make the client sessions. From log 1, we acquired 1738 extraordinary sessions and from log 2, 1987 sessions.

Day	No of Log Records (Log 1)	No of Log Records (Log 2)
1	17425	65536
2	16193	--
3	15214	--
4	11473	--
Total	60305	65536

Table 1: Raw network log entries

Hierarchical Levels	Data Objects (Sessions)	Did not participated
0	1985	0
1	1006	0
2	564	1
3	228	0
4	130	1
5	67	1
6	34	1
7	18	1
8	6	0
9	4	0
10	2	0
11	1	0

Table 2: Data set of session making

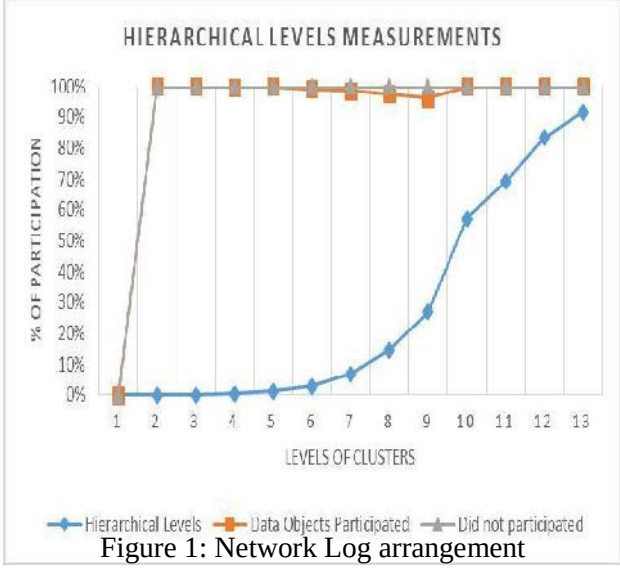

Figure 1: Network Log arrangement

Table 2 and Figure 1 are illustrations of chi-square based Assembling of network log. In Table 2 we have also mentioned the number of sessions, which did not participate in session pairing based on chi. Each image in Figure 3 represents the hierarchical clustering combination of the session at each level. For hierarchical Assembling, we take the 1985 sessions as independent clusters themselves. We compute the measurement for each cluster with the other clusters and paired the clusters that have maximum chi-square values. We marked the computation as level 1 and an average linkage criterion was applied for hierarchy generation. The same step was repeated for the generation of 2nd level hierarchy and so on. For this experiment we obtained 11 levels of hierarchy.

For the analysis of the proposed hierarchical clustering classifier, we used the precision and recall measures to evaluate the clustering results. We computed the true positive (TN), true negative (TN), false positive (FP) and false negative (FN) in each hierarchy level for the analysis of placements Figure 1: Hierarchy levels of network log Assembling of clusters in that particular level. The precision and recall results of 11 levels are shown in the Figures 2 and 3, respectively.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

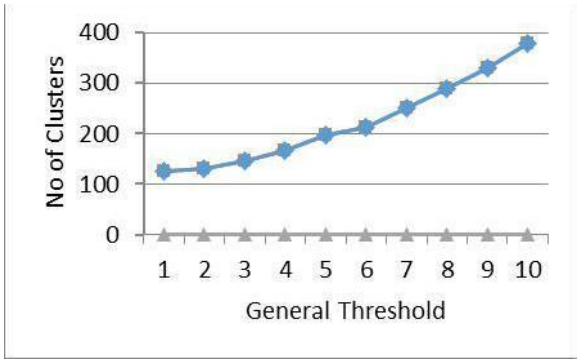

Figure 2: clusters used and General Threshold

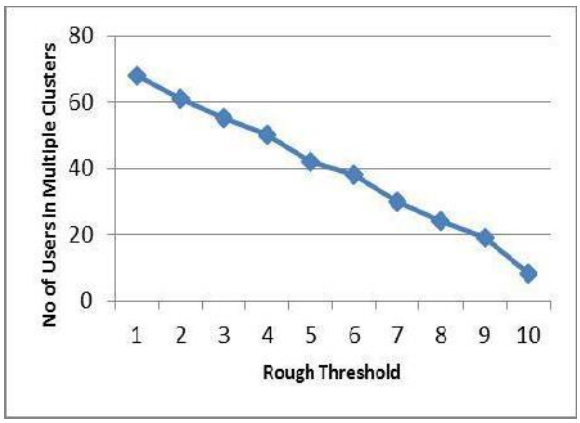

Figure3: No.of users Vs Threshold value

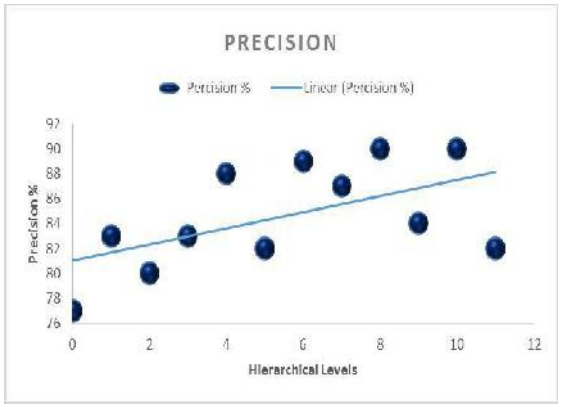

Figure 4: Average accuracy Level

VI. CONCLUSION AND DISCUSSIONS

The proposed method classifier is simple and effective to improve the visualization of network log. The results are verified on two other published classifiers. It helps to analyze the network log for predefined objectives. Number of pages and time spent by a single user in a session are the two parameters on which the measurement values are calculated.

REFERENCES

[1] P. Kolari and A. Joshi, "Network mining: research and practice," Computing in Science and Engineering, vol. 6, no. 4, pp. 49–53, 2004.

[2] W. Tong and H. Pi-lian, "Network log mining by an improved aprioriall algorithm," in Intl proceeding of world academy of science, engineering, and technology, pp. 97–100, 2005.

[3] B. Mobasher, "Data mining for network personalization",Lecture Notes in Computer Science, 4321:90, 2007.

[4] Wang Y.T. & Lee A.J.T., "Mining network navigation patterns with a path traversal graph", Expert Systems with Applications Vol.38 no.6, pp. 7112 – 7122, 2011.

[5] Vellingiri J., Kaliraj S., Satheeshkumar S. & Parthiban T., "A novel approach for user navigation pattern discovery and analysis for network usage mining", Journal of Computer Science vol.11 no.2,372 – 382, 2015.

[6] Chitraa V. & Davamani D.A.S., "A survey on preprocessing methods for network usage data",International Journal of Computer Science and Information Security Vol.7 no.3, pp. 78 – 83, 2010.

[7] Z. Ansari, M. Azeem, A.V. Babu, W. Ahmed, "A fuzzy approach for feature evaluation and dimensionality reduction to improve the quality of network usage mining results", Int. J. Adv. Sci. Eng. Inf. Technol. Vol.2 no.6 ,pp. 67–73, 2012.

[8] Z. Ansari, M.F. Azeem, A.V. Babu, W. Ahmed, "A fuzzy clustering based approach for mining usage profiles from network log data", Int. J. Comput. Inf. Sci. Secur. Vol.9 no.6, 70–79, 2011.

[9] A. Ketata, S. Mudur, N. Shiri, "Dependable performance analysis for fuzzy clustering of network usage data", IEEE Symposium on Computa-tional Intelligence and Data Mining, 2009, CIDM'09, pp.275–282, 2009.

[10] Park S., Suresh N.C. & Jeong B.K, "Sequence based clustering for network usage mining: a new experimental framework and ANN-enhanced K-means algorithm", Data and Knowledge Engineering vol.65 no.3, pp.512-543, 2008.