Information Technology Journal, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.

Improvement of Web Data Mining Method and its Application in Personalized Recommendation

Yao Chunlong, Sun Cuicui, Fan Fenglong and Shen Lan
School of Information Science and Engineering, Dalian Polytechnic University, Dalian, 110634, China

Abstract: Personality recommendation system in digital library has good development and applied prospects. It is becoming gradually an important research content in electronic resources intelligent processing. This study proposes a WEB data mining method based on the improvement of genetic algorithm. At the same time, this method is applied to the digital library electronic resources in personalized recommendation. The experimental results show that the method is suitable for large-scale text data set. This method in classification accuracy is higher and classification is faster. The method greatly improves the text mining system classification efficiency.

Key words: Data mining, personalized recommendation, genetic algorithm, digital library

INTRODUCTION

Knowledge Discovery in Database is refers to the whole process of found useful knowledge (Schapire and Singer, 2000). Data mining is a process of the particular steps. Knowledge Discovery in Database including data selection, pretreatment, data conversion, data mining, model to explain, knowledge evaluationand other steps (Sebastiani, 2002). It is a cycle iterative processes that explanatory model by the application of specific data mining algorithm. Itis finding the knowledge constantly deepening refinement, making it easy to understand. Data mining is a key step in the process of knowledge discovery. Data mining extract potential, unknown useful information, patterns and trends from a lot of, incomplete, noisy, fuzzy and random data. The purpose of the data mining is to improve the market decision ability, detecting anomaly pattern. These knowledge and rule is implicit, previously unknown but useful information for the decision. Through the data mining, valuable knowledge, rules or high level of information from the database and related data sets can be extracted for decision provides the basis, so that the database as a rich reliable resource for knowledge induction service.

The core technology of the data mining after more than ten years development has made great achievements. Today the data mining technology has entered a practical stage, because of the high performance relational database engine and the wide range of data integration (Bekkerman *et al*., 2011). WEB data mining data mining is an important branch, it development and put forward along with database technology, artificial intelligence technology and network technology.

The Internet makes the current digital library information resources richer but with the expansion of the information, there is difficult for user to obtain consistent with its preference feature information. In order to overcome this kind of difficult, personalized recommendation technology has been applied to the digital library. It can active referusers to its may need information. At present, the personalized recommendation technology (ZelIkovitz and Hirsh, 2011) has been applied in many fields, such as e-commerce, WEB information retrieval, etc. Which is relatively mature applicationis recommendation technology based on collaborative filtering. But there is many problems in its application process, such as the user appraisal matrix sparse solution algorithm, scalability, etc. In order to solve these problems, a lot of method has been put forward, such as single value decomposition method (Zeng *et al*., 2011), Bayes method (Parpinelli *et al*., 2011), etc. But the sparseness have not been very good solved. The main reason is that reader's interest is associated with professional background.

For most readers especially research readers, their interests focus primarily on one (or some) field. They are very interested in the information of research field and lose interest in outside. In view of this, this study proposes a genetic algorithm based on the improvement of WEB data mining method. At the same time, this method is applied to the digital library electronic resources in personalized recommendation.

Corresponding Author: Yao Chunlong, School of Information Science and Engineering, Dalian Polytechnic University, Dalian, 110634, China

MATERIALS AND METHODS

Technology of text classification: Text classification technology appeared in the early 1960's (Yuan *et al*., 2011). In the 1980's, the researchers used Knowledge Engineering (KE) method to realize the text classification. KE method usually uses Disjunctive normal Form for each class definition logic rules (Hu and Hu, 2011). It is a kind of simple Natural Language Processing method. Rau adopting vocabulary complicated to realize the Natural Language for Data Bases classification. Jacobs adopting statistical methods to assist structure classification rule, further improve the performance of text categorization system (Wang and Wang, 2005). The text categorization KE method need to manually compilation rules or application other complicated NLP techniques, the difficulty is very big and time-consuming, too low and not practical on many occasions.

Web mining: WEB mining is the process discovery and extraction information automatically by data mining technologyfrom WEB documents and service. WEB mining is a comprehensive technology, involving WEB, data mining and computational linguistics, informatics and other fields (Yang, 1999). Compared with the traditional data mining, WEB mining has a lot of unique (Yang, 2010). First of all, WEB is a document node and links from a graph in logic, so the mode by WEB mining can be about WEB content, also can be about the structure of the WEB. Secondly, WEB mining object is large, distributed, heterogeneous WEB document, they are structured and semi-structured data quantity, large, rapid growth and has the semantic machine difficult to understand (Yang and Pedersen, 1997). So the existing data mining tools is not completely suitable for WEB mining. In this way, the development of new WEB mining technologyhas become the focus of WEB mining research.

Web mining steps: At present, according to the general data mining method and in combination with the characteristics of Web data, Web data mining is divided into the following five steps, as shown in Fig. 1:

- **Data sampling:** Web environment at present can provide data sources including Web page data, data links and log data records of the user access, etc. According to the principle of topics, data sampling from a large number of data take out data subset target related, for the data mining provides materials and resources
- **Data preprocessing:** Data preprocessing is the processes and organization reconstruction for data, constructing the topics of data warehouse and provide a basic platform for further data mining process. It mainly includes: Data cleaning, data integration, data transformation and data contracted
- **Data mining:** This is the core part of the data mining system. Its main function is to use all kinds of data mining technology, derived the potential and effective and can be understood the knowledge model from data after pretreatment. The goal of data mining is to describe and forecast. Descriptive model is decryptingthe rules existing in data, or clustering data according to the similarity of data. The prediction is finding out its regularity based on the attribute refers to the existing data value and speculate on the future possible attribute value
- **Analysis and assessment:** Data mining the knowledge model need reliability and validity analysis and evaluate its conclusion. Provide information support for the management and decision-making of the user. How to check for the results of the analysis is useful, a simple method is to directly use the original model and the data of the sample inspection. Another way is to find another some actual data reflect the objective regularity to test. The third way is take out new data in the actual operation of the environment to test
- **Knowledge representation:** Knowledge representation is demonstratingthe knowledge

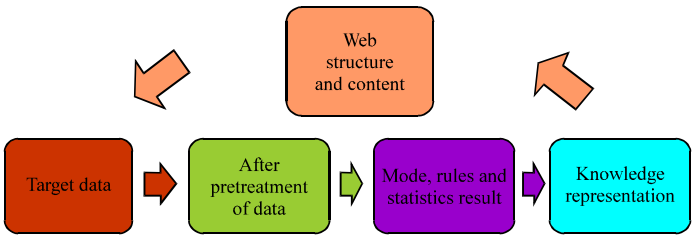

Fig. 1: Algorithm process of Web data mining

modemined out from Web data use the data mining tools in the appropriate form, for users to accept and communicate with each other. The task of data mining is various, mainly including: To summarize Association rules mining, Classification rule mining, Clustering rules mining, Prediction analysis, Trend analysis, Deviation analysis, etc.

IMPROVEMENT OF WEB DATA MINING METHOD

Improvement of WEB data mining method: The optimization framework of WEB data mining methods improvedbased on the genetic algorithm is shown in Fig. 2. The method according to the order of the covered way, trying to dig out a classification rule list covered most or the entire training sample. This method can be used to describe the implementation process as follows:

- **Variable initialization:** Set rules list have found empty. At the same time, the entire training sample placed in the training sample concentration
- **Evolution of the genetic algorithm:** Genetic algorithm every time evolution can find a classification rule. After evolution of genetic algorithm, the rules joined to the list of rules that have been found. At the same time, the rules cover sample eliminated from the training sample concentration
- **Termination conditions:** When the number of samples did not cover is less than the user preset value, the genetic algorithm stop evolution

Chinese WEB text mining prototype system by use of this method is not only an experiment platform but also a practical platform. On the platform, it can be WEB text mining process of vector representation, feature extraction, classification and experiment and can be

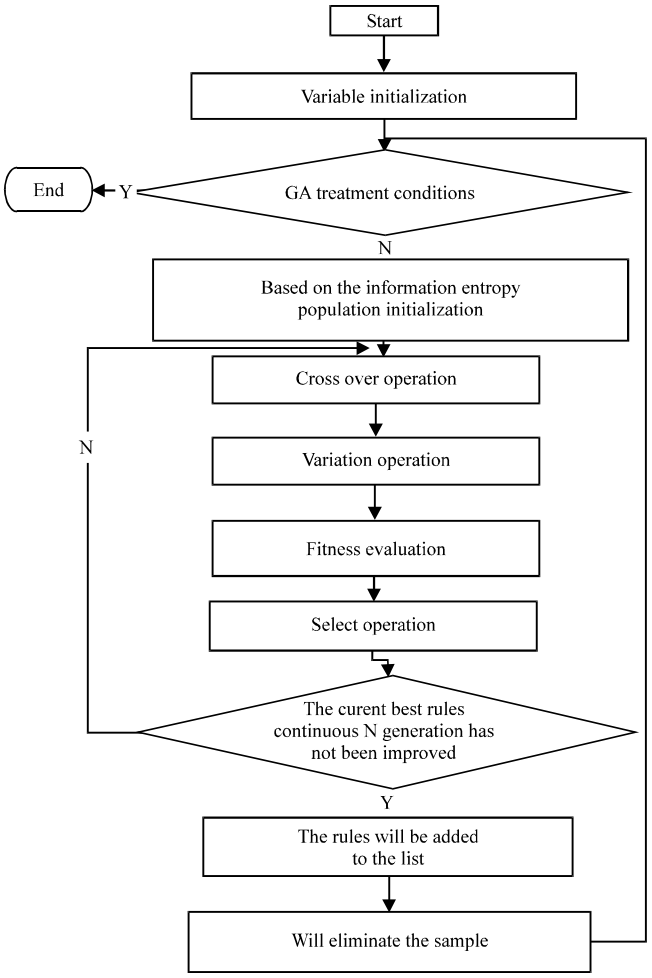

Fig. 2: Framework of improve WEB data mining methods optimization

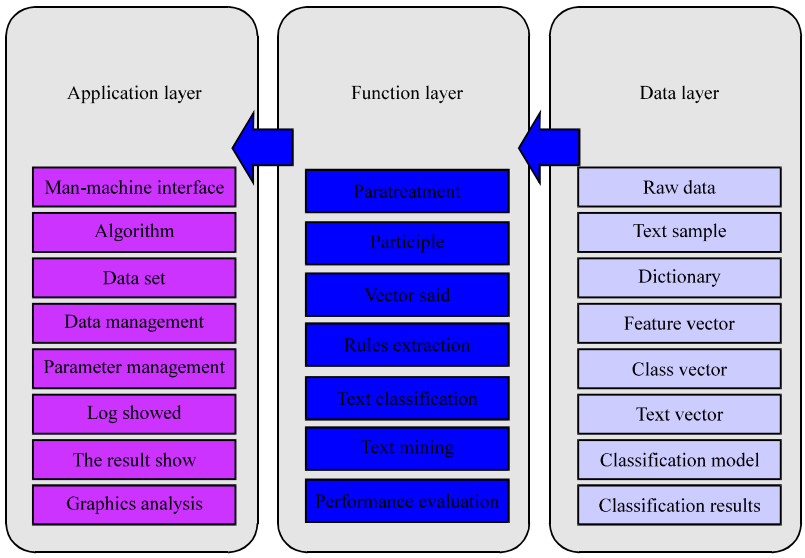

Fig. 3: Chinese WEB text mining prototype system level structures

collected on the original WEB page according to predefined category select classifier for automatic classification. As a practical platform, the independence of corpus and algorithm are considered before the system design. The independence of corpus means the mining object can be original WEB page or plain text files. The independence of algorithm means mining processes of the key link is relatively independent. Through setting various static data, different algorithms, parameters and data file can be choosing in the classification of each stage. At the same time, getting optimization algorithm formation parameter set in the experimental stage, so that the result of the experiment can be uses directly for the use in the actual. Using the thought of object-oriented program design, our system can be divided into three layers from the structure, as shown in Fig. 3:

- **Data layer:** Encapsulate the different kinds of data in the system and the basic operation, into class, all of the data using basic unified data structure. The data continuous index and needs often inquires using hash table way, using unidirectional list of ways to solve problems that "conflict" for other data in the hash table internal. For other data use the structure array way storage. On the one hand, to consider the data reading and writing, query efficiency, on the other hand consider the decrease as far as possible the use of memory
- **Function layer:** Realize all module function on the basis of data layer. Functional modules based on the operation of the data type, to realize the basic function of the system, complete the various testing and automatic classification process
- **Application layer:** Complete man-machine interface, data organization, management of the parameters and recordthe various test and classification process of processing steps. Realize outputting the process log file and displaying the analysis results of the graphical. Through the analysis dependence and inheritance of various data, guarantee the consistency of the data mining process

Personalized recommendation algorithm based on association rule: The main goal of personalized recommendation algorithm is generating the recommended matching sets according to the current user access operation and user Settings of the recommended parameters. Recommend is set by the user access operation of the web page.

The core of the algorithm is using a fixed size sliding window covering the current user access operation sequence. It's a effective method to realize online personalized recommendation service. The sequence within sliding window constantly updated with the progress of the access process. For example, the size of the sliding window is 3, the user access operation sequence within the sliding window is <A, B, C>. When users access to the D, the new sequence in sliding window is <A, B, C, D>. Under normal circumstances, if the sliding window size is n, so only the recent visit by n web page affect recommended collection. So it is meaningfulto personalized recommendation service. Because if the length of the current user access operation sequence is too big, it is difficult to obtain so much

Table 1: An example of largest forward access paths set meeting minimum support degree

Serial No.	Sequence	Support
1	A-B-C-D-E-F	11
2	A-G-C-D-H-F	10
3	O-A-B-C-D-E	9
4	O-A-B-C-D-F	9
5	A-I-C-D-J-F	7
6	O-A-I-C-D-J	7

information, namely a few of the match, not even withitems matching operation. With a short sequences will have too much of the match, the recommended service is also bad. Therefore, the influence of sliding window size on recommended set is a problem worthy of research. Existing experiments show that the average length of the visit as a sliding window size is better choice. Table 1 give an example of largest forward access paths set meeting minimum support degree.

From a new point of view, recommend a physical link from the user access operation of the web page is preferred. Therefore, a link distance factor defined for selection strategy. Physical link path length determined by site topology directed graph. Each node in digraph represents of a URL of web page in the site. If there is a physical link from web page X to web page Y, then the corresponding node X to Y exist a edge. The physical link path distance between two URL (i.e., u1 and u2) is defined as: The minimum path length of the visit from u1 to u2 in site digraph.

Given gathered tree directed graph G = (V, E), the V⊂U. "s" is the current user access operating sequence, u is a URL from the current recommended and u ∉s. Dist (u, s, G) said the URL of the smallest physical link path distance between u to s, link distance factor calculation equation is:

$$idf(u, s) = \log(dist(u, s, G))+1$$

If u belongs to the current user's access operation sequence, the definition of link distance factor is 0.

Business model of pretreatment for the users is based on the support filter method. After this step processing, a lot of don't frequent user transaction mode is flit out, reduce the dimension of the features of the user affairs. The vector of transaction mode characteristic is using the simple binary representation. Assume T is a transaction mode set with filter method. It's expressed as T = {t1,..., tm}, ti = < s1,... Sn > and ti ∉T. N is the number of different URL in T. M is the number of user transaction modes in T. Sj represents the value of a visit URL_j corresponding operation. It expressed as:

$$S_j = \begin{cases} 1, URL_j \in t_i \\ 0, otherwise \end{cases} \quad (1)$$

For t_i and t_j transaction model, This study use the cosine measure the similarity of it:

$$S_j = \begin{cases} 1, URL_j \in t_i \\ 0, otherwise \end{cases} \quad (2)$$

This study use the recursion clustering algorithm in dynamic hierarchy index tree to realize the cluster of user transaction mode.

For the convenience of the recommendation, this study need to reduce the number of class URL. First each user transactions clustering model expressed as a weighted vector dimension of p, p is the number of different URL. Clustering center vector expressed as: Ci = {Wi..., Wp}, the weights for each URL is the class frequency. If n_j expressed as the frequency of URL_j in class i. Ni expressed as the number of user transaction mode in class i. URL_j weights expressed as:

$$w_j^i = w(URL_j, C_i)\begin{cases} \frac{n_j^i}{N_i}, URL_j = C_i \\ 0, otherwise \end{cases} \quad (3)$$

Next, filter URL with lower support and converts user affairs clustering model in URL clustering model. E is sited as a threshold epsilon, URLj in user transactions clustering model meet URL clustering model of :

$$SIM(t_i, t_j) = \frac{\sum_{K=0}^{N} S_k^i S_k^j}{\sqrt{\sum_{k=0}^{N}(S_k^i)^2}\sqrt{\sum_{K=0}^{N}(S_k^j)^2}}$$

For example, if u = 0.5, each URL in class with at least 50% present in the transaction mode. When URL clustering pattern form, the URL clustering model which match the URL clustering model of the current user access operation, as a candidate recommendation.

This section mainly discusses using the URL clustering the method to realize the recommendation service. Suppose s for the user's current visit operation path, s = <s1,... Sp>, if the user visited URL_j, s_j = 1, otherwise the sj = 0. In current access operation path and URL to calculate similarity, similarity computation equation is as follows:

$$Match(s, C_i) = \frac{\sum_{j=1}^{p} w_j^i S_j}{\sqrt{\sum_{j=1}^{p}(w_j^i)^2}\sqrt{\sum_{j=1}^{p}(S_j)^2}} \quad (4)$$

where, T is the minimum matching threshold used to determine whether match, to determine the minimum

threshold method according to the statistical situation of access Log files. If found the URL matching clustering (may be match more than one URL clustering), then it need to evaluate the URL in each of the URL matching clustering and choose the right URL as recommended.

RESULTS AND ANALYSIS

Test of method: In order to validate the efficiency of the method proposed in this study the author usesthree standard text data sets. (1) 20 newsgroups (20 NC). Twenty NC is a common text data set, which collected from twenty newsgroup nearly 20000 news. Twenty NC common contains 18828 texts. (2) Industry Sector (IS). IS is a web page data set, the site were from various industrial economy department website. This data set contains 9652 pages in total text, belongs to 105 different classes. (3) Web data set (Web). Web collected from Google's Open Directory Project project. In the experiment, this study select randomly 35 classes, a total of 5035 page as experimental data set.

The author has used two kinds of typical test method to evaluate this study WEB text mining system performance. One is training-testmethod and the other is k-fold cross validation method. Training-test method is classic evaluation method, it will initial sample set into training set and testing set two parts. Training set used for feature selection and classifier training. Test set used to test classifier. k heavy cross check will divide samples into k pieces, every time take the k-1 as a training set. The rest as a test set and take their average value as the final result. The author select recall ratio and precision ratio as evaluation this study WEB text mining system two performance indicators.

Computer configuration is Pentium IV processor, basic frequency for 214 g, memory for 512 m to complete this test. In order to avoid randomness in the process, each experiment was repeated twenty times. Take the average of twenty times as the final experimental results. Test results for three text data sets by use of the WEB text mining system is shown in Fig. 4. The results in recall ratio of 20 nc 'S' WEB income out of WEB text mining system are ideal.

Application of the proposed method: Experimental data is access data that readers in the university digital library access to the resource pool. In the experiment, this studyuse the implicit way (such as skimming, download, etc.,) for readers to read in a library of the level of interest, to avoid the extra burden that require the readers to evaluate the explicit feedback.If the reader not interested in one book and no corresponding evaluation feedback information, it will lead to final test data not accurate. At the same time the timing function joined in system, the time of reader read one book often can reflect interest level of this book. In the experiment, this studyanalyzed each evaluation level of the corresponding reading time and then the system can get the evaluation information of readers to books in a library. For download e-book is set to the highest rank of this interest.

WEB server LOG has a complete structure, visit the page, time, user ID and other information, has a corresponding record in the LOG when users access to WEB sites. Hot WEB site can record hundreds of megabytes WEB LOG records every day. Use mining tools to mining the log file. This study can learn the user's access model and improve intelligent WEB site information organization and display according to the specific user. For WEB site managers, high efficiency automatic aided design tool can improve their work efficiency. The tools can dynamic adjust the organization and display on the WEB site according to the visitor's interest and access time. For users, they hope to use intelligent tool to find desired information resources, tracking and analysis their browse mode. Hope to see the personalized page and get better satisfy service.

The personalized recommendation system mainly includes three modules: Data preprocessing module, pattern mining module and mode analysis module. In data preprocessing module, filet the "unk" data of the WEB

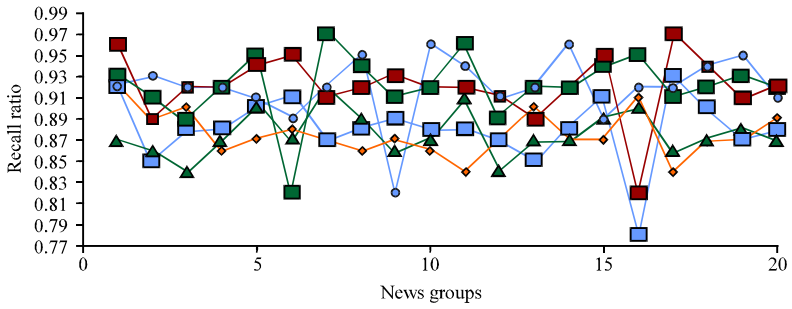

Fig. 4: Results of WEB text mining system performance evaluation

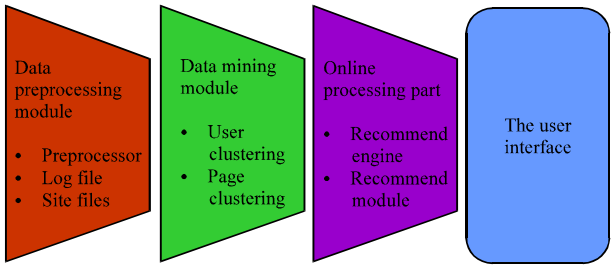

Fig. 5: Framework of electronic resources personalized recommendation system in digital library

Server AccessLOG file, through the original data collection, cleaning, conversion, reduction of several steps. Make the purification data into the form of a transaction database, in order to convenient pattern mining stage use. In the pattern mining module, get the page and user groups of clustering results through things database analysis, so as to understand the web page classification and user interest trend. Mode analysis and application modules, mainly is the application of weight matrix clustering algorithm the page and user groups of clustering results. Electronic resources personalized recommendation system in digital library design as shown in Fig. 5.

Due to the personalized recommendation is the effective methodto improve the network efficiency and attract users on the network to access. So in today's information society which website as information node, the system will have broad application prospects and great practical significance.This study believe that the user oriented personalized recommendation will be more perfect, along with the WEB log data analysis and research is unceasingly thorough.

CONCLUSION

The main innovation points: The study presents an improvement WEB data mining method based on the genetic algorithm. At the same time, this method is applied to the electronic resources in personalized recommendation in digital library. The experimental results show that the method is suitable for large-scale text data set. The method extracting rules of the classification accuracy is higher, classification fast. The method greatly improves the text mining system classification efficiency.

REFERENCES

Bekkerman, R., R. El-Yaniv, N. Tishby and Y. Winter, 2011. On Feature Distributional Clustering for Text Categorization. ACM Press, New Orleans, Louisiana, pp: 146-153.

Hu, Z.J. and M.S. Hu, 2011. Multi agent web text mining system based on grid. Microcomput. Inform., 26: 266-268.

Parpinelli, R.S., H.S. Lopes and A.A. Freitas, 2011. A data mining with an ant colony optimizations algorithm. IEEE Trans. Evol. Comput., 6: 321-332.

Schapire, R. and Y. Singer, 2000. BoosTexter: A boosting-based system for text categorization. Mach. Learn., 39: 135-168.

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Comput. Surveys, 34: 1-47.

Wang, Y. and Z.O. Wang, 2005. Text categorization rule extraction based on fuzzy decision tree. Comput. Appl., 25: 1634-1637.

Yang, Y.M. and J. Pedersen, 1997. A comparative study on feature selection in text categorization. Proceedings of the 14th International Conference on Machine Learning, July 8-12, 1997, Nashville, TN., USA., pp: 412-420.

Yang, Y., 1999. An evaluation of statistical approaches to text categorization. Inform. Retrieval, 1: 69-90.

Yang, Y.M., 2010. An evaluation of statistical approaches to text categorization. Inform. Retrieval, 26: 66-68.

Yuan, H.B., A. Li and B.Q. Hu, 2011. Method of web content mining based on XML. Microcomput. Inform., 26: 196-197.

Zellkovitz, S. and H. Hirsh, 2011. Using LSI for text classification in the presence of background text. Proceedings of the 10th International Conference on Information and Knowledge Management, (IKM'11), ACM Press, New York, pp: 113-118.

Zeng, H.J., Z. Chen and W. Yma, 2011. A unified framework for clustering heterogeneous web objects. Proceedings of the 3rd International Conference on Web Information Systems Engineering, December 12-14, 2011, Washington, DC, USA., pp: 161-170.