International Journal of Soft Computing 12 (5-6): 303-307, 2017
ISSN: 1816-9503

An Improved Web Emotion Analysis using Hybrid PAM Neural Network Approach

Meera Alphy and Ajay Sharma
Department of Computer Science and Engineering, SRM University, Delhi-NCR, Sonepat, Haryana

Abstract: Web usage mining method for learning episodic web access patterns from web usage logs which integrates knowledge on customer's interest and behaviours. A novel method to efficiently provide better web-page recommendation through semantic-enhancement by integrating the domain and web usage knowledge of a website is introduced in this study. Here, we propose, a hybrid approach by first enriching the contextual information by k-medoids algorithm and training each cluster using simple Neural Network approach. This method improves the cluster quality in term of accuracy and CPU time when compared to traditional clustering techniques like k-means clustering and deterministic genetic clustering techniques.

Key words: CPU, access patterns, clustering techniques, integrating, efficiently, accuracy

INTRODUCTION

Data mining is digging the useful data from the large set of data and using different methods like artificial intelligence, machine learning and statistics from the database systems. Web mining is the application of data mining techniques to unearth motif from the world wide web. Web mining can be subdivided into web usage mining, web content mining and web structure mining. This subdivision in basics of:

- Web user activity can be get from file saved in server and online user's browser tracking
- Web graph can be get from links between pages, people and other data
- Web content can be achieved for the text or data found on web pages and inside of documents

Web usage mining defined to the automatic unearthing and testing of patterns of online users and related data collected as a output of user communication with web resources on the different or related web sites. The aim is to get, arrange and relate the behavioural patterns and profiles of users communicating with a web site. The unearthed patterns are usually symbolized as collections of pages, objects or resources that are more often accessed by groups of users with familiar needs or interests. Similarly, standard data mining process, the web usage mining process can be divided into three stages. They are data collection and pre-processing, pattern discovery and pattern analysis. In the pre-processing stage, the data is distinguish useful information from raw data and divided into a set of user communication representing the behaviour of each user during different visits to the site.

Applications of web usage mining are business intelligence, competitive intelligence, pricing analysis, events, product data, popularity, reputation. Recommended systems are a subclass of information filtering system that hunt for to foresee the rating or preference that a online user would give to an item.

Recommender systems characteristically develop a list of recommendations in any of these two ways mentioned below. Collaborative filtering approaches create a method from a user's past behaviour as well as similar decisions made by other users. This method is then used to predict items that the user may have similar interest in. Advantage of this method is no knowledge-engineering effort, serendipity of results, learns market segments. Disadvantage of this method is requires some form of rating feedback, cold start for new users and new items content-based filtering method uses a series of different characteristics of an item in order to recommend additional items with similar properties. Advantages of this method are no community required, comparison between items possible. Disadvantages are content descriptions necessary, cold start for new users, no surprises. Knowledge based recommender systems are products with low number of ratings. Advantage of this method is deterministic recommendations, assured quality, no cold-start can resemble sales dialogue. Disadvantage of this method is knowledge engineering effort to bootstrap, basically static does not react to short-term trends.

Corresponding Author: Meera Alphy, Department of Computer Science and Engineering, SRM University, Delhi-NCR, Sonepat, Haryana

Literature review: Fong *et al.* (2012) introduced a semantic web usage mining approach for find out periodic web access patterns from noting web usage logs through self-reporting and behavioural tracking based on not related information on consumer emotions and behaviours. Personal web usage lattice that models the user's web access activities are incorporate with behavioural and emotional cues, represents fuzzy temporal and resource. On this output researcher generate a personal web usage ontology which enables personalized web resources recommendation. Nguyen *et al.* (2014) introduced a novel method to efficiently provide better web-page recommendation through semantic-enhancement by integrating the domain and web usage knowledge of a website. Two new models are put forwarded to represent the domain knowledge. They are:

- The first model uses ontology to correspond to the domain knowledge
- The second model uses one automatically generated semantic network to correspond to the domain terms, web-pages and the relations between them
- Another new model, the conceptual prediction model is automatically developed a semantic network of the semantic web usage knowledge

Zhu *et al.* (2014) proposed an approach for first enriching the contextual information of mobile apps by exploiting the additional web knowledge from the web search engine. All the enriched contextual information into the maximum entropy model for training a mobile app classifier. To validate the proposed method, an extensive experiments on 443 mobile user's device logs to show both the effectiveness and efficiency of the proposed approach. Tang and Liu (2014) investigated a novel problem of feature selection for social media data in an unsupervised scenario by proposng a novel unsupervised feature selection framework, LUFS for linked social media data. We systematically design and conduct systemic experiments to evaluate the proposed framework on data sets from real-world social media websites. Geng and Tian (2015) a new method to identify navigation related web usability problems based on comparing actual and anticipated usage patterns. The actual usage patterns can be extracted from web server logs routinely recorded for operational websites. The anticipated usage including information about both the path and time required for user-oriented tasks is captured by our ideal user interactive path models constructed by cognitive experts based on their cognition of user behaviour. A software tool was developed to automate a significant part of the activities involved. Kang *et al.* (2016), a new web service recommendation method with diversity feature of user interests on web services and reduced a user's potential quality of service used. Web services are first unearthed by detail study on the web service usage history based on web user's interests and quality of service preferences. Historical and potential user interests and their QoS utility are factors helped in indexing of web service candidates. And also developed a web service graph in the basis of functional similarity between web services. In the last phase, researcher constructed diversity-aware web service ranking algorithm to rank the web service user's depends on their scores and diversity degrees which was obtained from the web service graph. Liu *et al.* (2016) the Request Dependency Graph (RDG) which described the relationships among HTTP requests through a graph. The contacting of web object B is caused by the contacting of A which represent a directed link from A-B in the graph. Researcher proposed a method to create such a graph by taking out the temporal and causal information among grouped HTTP requests. To exhibit the value and effectiveness of the proposed model, researcher designed an algorithm for primary requests identification based on the RDG. The k-means algorithm is partitioning an N-dimensional population into k sets (Macqueen, 1967). In algorithm each cluster is related with a cluster centroid where each data point is allocated cluster having closest centroid.

MATERIALS AND METHODS

Proposed system: Dataset is a gathering of related sets of information that is collected by different elements but can be manoeuvred as a unit of computer. Web log file stored and update by a server consisting of a list of task user performed. Pre-processing is done to identify useful data from raw dataset. Figure 1 show proposed PAM-NN algorithm.

In Fig. 1, pre-processing is done to identify useful data from raw dataset. The pre-processed data analyze and check the correlation between data's defines the pattern discovery which here done by hybrid algorithm using PAM and NN. Entropy, purity and DB index are

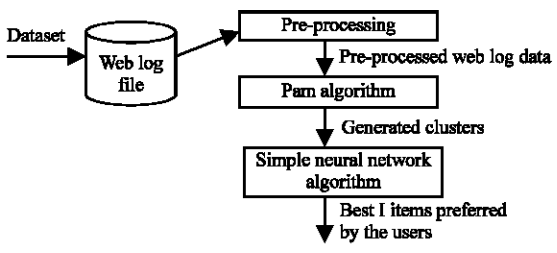

Fig. 1: PAM-NN architecture

used for pattern analysis (validation and interpretation of the mined patterns). Proposed PAM-NN system show in Algorithm 1.

Algorithm 1 (PAM-NN algorithm):
Input: dataset D° Rn, where |D| = d
Output: best I items preferred by the users
1. Read the data items
2. Run PAM algorithm
3. Run simple-NN algorithm
4. End

K-medoids is a a partitioning algorithm that divides dataset up into sets. It decrements the sum of variation among data items that are the members of a cluster and the a data item as the centre of that cluster. K-medoids separates data into k clusters with k known a priori. A data item whose average dissimilarity to all other data items is minimal is taken as medoid. It is the most centrally located point in the set. The most centrally located point in the set. The most common realisation of k-medoid clustering is the Partitioning Around Medoids (PAM) Algorithm 2 as follows.

Algorthm 2 (PAM algorithm):
Notations D = dataset; d = number of data items; R = n dimensional real number space.
Input: dataset D° Rn, where |D|=d
Output: User profiles, Clusters C1 to Cn
Steps:
1. Repeat
2. randomly select k of the n data points as the medoids
3. Associate each data point to the closest medoid
4. For each medoid j and each data point o associated to m swap m and o and compute the total cost of the configuration using

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, m_i)^2$$

5. Select the medoid o with the lowest cost of the configuration
6. Repeat alternating steps 3 and 4 until there is no change in the assignments

The user profiles generated from the clusters obtained by the PAM algorithm are trained using simple neural network technique as explained in Algorithm 3.

Algorithm 3 (Single Layer Perceptron NN algorithm):
Input: User profiles
Output: best I items preferred by the users
Notations used:
T = desired output, O = output generated by the perceptron, w_i = weight associated with the ith connection
Initialize weights w_i set to small random values
Learning rate $\lambda = 0.1$
Repeat
For each training example (xa, T)
Calculate output activation $O = f\left(\sum_{i=o}^{d} w_i x_{ia}\right)$
If (O not equal to T) then
Update the weights $w_i = w_i + \lambda (T\text{-}oa) x_{ia}$
End if
End For
Until weights don't change

A neural network is a machine learning approach motivated from the brain performs a particular learning activity:

- Information about the learning activity trailed in form of examples
- To deposit the acquired information, inter neuron link weights are used
- The weights are customized in sort to form the particular learning activity correctly for the period of the learning course of action on the training examples

Algorithm for simple perception performance as input is perceptron which has random weights and a training set and perception output does not match the actual output and then change the weight by an amount proportional to the difference between the desired output and the actual output. Equation for preceptron learning rule is given in Eq. 1:

$$\Delta W_x = n \times (out_F - out_B) I_x \quad (1)$$

Where:
n = The learning rate
out_F = The desired output
out_B = The actual output

These kind of recommendations gives higher quality recommendations by effectively understanding users interest and emotions.

RESULTS AND DISCUSSION

Evaluation of results: The entropy (Tan *et al*., 2005) is calculated using the following equation:

$$e_x = \sum_{x=1}^{L} p_{xy} \log_2 p_{xy} \quad (2)$$

$$p_{xy} = \frac{m_{xy}}{m_y} \quad (3)$$

Where:
L = Number of classes
p_{xy} = Probability that a member of cluster y belongs to class x
m_y = Number of values in cluster y
m_{xy} = Number of values x in cluster y

The overall entropy of a clustering approach is given by:

$$e = \sum_{x=1}^{k} \frac{m_y}{m} e_y \quad (4)$$

Where:
m_y = Size of the cluster y
k = Number of clusters
m = Total number of data points

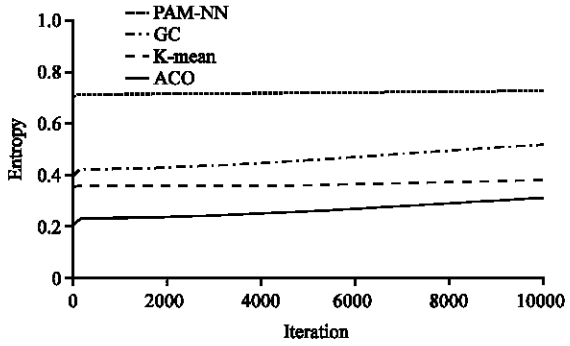

Fig. 2: Quality measure in terms of entropy

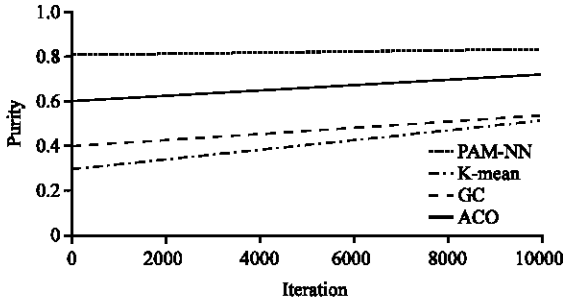

Fig. 3: Quality measure in terms of purity

From the Fig. 2, we may observe that our proposed method has higher value of entropy than traditional approaches like k-means, genetic algorithm and ant colony optimization. Purity (Tan *et al*., 2005) is calculated using the equation:

$$\text{Purity}_m = \max p_{mn} \tag{5}$$

The overall purity of a clustering approach is given by:

$$\text{Purity} = \sum_{i=1}^{j} \frac{a_m}{a} \text{purity}_m \tag{6}$$

From Fig. 3, we may observe that our proposed method has higher value of purity than traditional approaches like k-means, genetic algorithm and ant colony optimization. Davies-Bouldin (DB) Index (Davies and Bouldin, 1979) calculates the average similarity between each cluster and its most similar one in a clustering result. DB index is defined by:

$$DB_{xy} = \frac{1}{x_y} \sum_{z=1}^{x_y} S_z \tag{7}$$

Where:

S_z = $\max_{z=1,\ldots,xy,\ z\neq y} S_{zw}$, x = 1, ..., x_y

x_y = The number of clusters

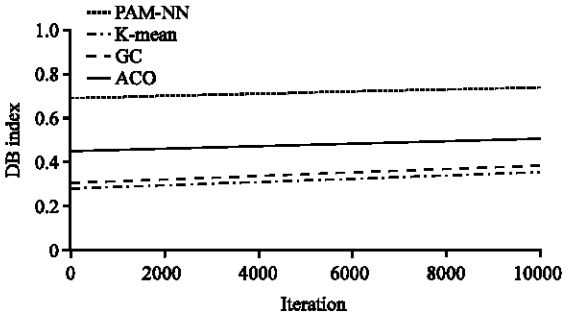

Fig. 4: Quality measure in terms of DB index

From Fig. 4, we may observe that our proposed method has higher value of DB index measure than traditional approaches like k-means, genetic algorithm and ant colony optimisation.

CONCLUSION

In this study, a hybrid approach for web emotional analysis is proposed. The proposed system is initial clusters are obtained using k-medoid algorithm and the patterns in each clusters are trained using simple neural network approach to analyse the emotions of web users. The experiment result shows that when compared to traditional algorithms like k-means clustering algorithm, genetic clustering algorithm and ant colony optimisation algorithm. Our proposed method gives better quality results in terms of entropy, purity and DB indexed. In this research limited parameters are used. In future, this research can be extended to user tracking profiles.

REFERENCES

Davies, D.L. and D.W. Bouldin, 1979. A cluster separation measure. IEEE. Trans. Pattern Anal. Mach. Intel., 1: 224-227.

Fong, A.C.M., B. Zhou, S. Hui, J. Tang and G. Hong, 2012. Generation of personalized ontology based on consumer emotion and behavior analysis. IEEE. Trans. Affective Comput., 3: 152-164.

Geng, R. and J. Tian, 2015. Improving web navigation usability by comparing actual and anticipated usage. IEEE Trans. Hum. Mach. Syst., 45: 84-94.

Kang, G., M. Tang, J. Liu, X.F. Liu and B. Cao, 2016. Diversifying web service recommendation results via exploring service usage history. IEEE. Trans. Serv. Comput., 9: 566-579.

Liu, J., C. Fang and N. Ansari, 2016. Request dependency graph: A model for web usage mining in large-scale web of things. IEEE. Internet Things J., 3: 598-608.

Macqueen, J., 1967. Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, Le Cam, L.M. and J. Neyman (Eds.). University of California Press, Berkeley, CA., USA., pp: 281-297.

Nguyen, T.T.S., H.Y. Lu and J. Lu, 2014. Web-page recommendation based on web usage and domain knowledge. IEEE. Trans. Knowl. Data Eng., 26: 2574-2587.

Tan, P.N., M. Steinbach and V. Kumar, 2005. Introduction to Data Mining. 1st Edn., Addison Wesley, New York, ISBN-10: 0321321367, Pages: 769.

Tang, J. and H. Liu, 2014. An unsupervised feature selection framework for social media data. IEEE. Trans. Knowledge Data Eng., 26: 2914-2927.

Zhu, H., E. Chen, H. Xiong, H. Cao and J. Tian, 2014. Mobile app classification with enriched contextual information. IEEE. Trans. Mobile Comput., 13: 1550-1563.