Implementing Cluster Analysis on An E-Learning Website

Neha Goel*

Vivekananda Journal of Research
July- Dec 2019, Vol. 8, Issue 2, 158-170
ISSN 2319-8702(Print)
ISSN 2456-7574(Online)

http://www.vips.edu/vjr.php

Abstract

Data is increasing every day. But this data needs to be converted to information so that it can be utilized for strategic decision making. Data Mining is one such area which helps in churning out useful information from the abundant data available. It is the extraction of interesting patterns or knowledge from huge amount of data. Mining is gaining huge popularity these days. To perform data mining a majority of techniques exist. Clustering is a technique of mining which works on the principle of understanding the differences and similarities between the data. In order to understand the effectiveness of an e-learning website, clustering technique was applied on the data extracted from the feedback forms. Feedback forms highlighted some questions which would help the developer in enhancing the features on the website. The data extracted was analyzed using MS Excel Add-in for Data Mining. The results and interpretations obtained have been discussed. The same can be helpful in attracting more users to the website.

Keywords- *Users, E-learning, Cluster Analysis, Website.*

INTRODUCTION

Data mining is referred to as Knowledge Discovery in Databases (KDD). It extracts information hidden in the database which can either be related to machine learning or exploratory data analysis that is being widely used these days. Some of the areas include prediction and description, relationship marketing, customer profiling, outlier identification and detecting fraud, customer segmentation, website design and promotion. The techniques of data mining include Classification, Clustering, Regression, Summarization and Link

* Assistant Professor, Vivekananda Institute of Professional Studies, New Delhi.
E-mail : nehagoel123@gmail.com

Analysis. Data mining can be applied on any type of data. It can be textual data, multimedia data, and data from the World Wide Web or any other source. Mining starts with data collection followed by Preprocessing, Pattern Discovery and finally Pattern Analysis[1].

Mining if performed on data from WWW is called Web Mining. The area of Web mining is gaining huge popularity these days as the web generates millions and trillions of data every day. Web Mining can be classified into three types of mining namely Content which deals with the Content data present on the web), Structure (which examines the Hyperlink structure of web) and Usage dealing with the data stored in Web access logs of server) Mining [2].The only way to utilize such data is by deploying data mining techniques on it.

In the current study an e-learning website was taken and in order to judge the response that how effectively the website has been designed Cluster Analysis was carried out. The e-learning concept ensures that the learning material is available online and also facilitates the mechanism of feedback. Feedback from the students help in ensuring that the course material provided is relevant and is able to satisfy their queries completely. Web based, computer based, virtual classroom and digital collaborations are some applications of E-learning. This increases the power of traditional text book materials when combined with online resources. Students get a good feel sitting at their home looking at rich media and interactive content. On sitting in this virtual classroom one can have freedom of covering the topics which were skipped initially. It helps in raising the level of education and literacy.

Clustering is a technique in data mining that assembles data of similar types in clusters. It is a form of unsupervised learning. The clusters are formed for a group of classes which are unknown and hence, it is considered to be a part of unsupervised learning. The major objective of clustering is to minimize the interclass similarity and maximize the intra class similarity depending upon some criteria defined on the object of attributes.

The paper has been organized as follows. Next section discusses the work carried out in the related area. Section 3 presents the experimental setting. Section 4 states the results and discussions obtained after implementing clustering on the available data sets. The conclusion and future scope of the same have been discussed in the end.

LITERATURE REVIEW

As stated previously web personalization has been adopted in almost all the areas

of WWW. Due to increase in the market competition, website owners are adopting to new strategies so that they can attract more and more users. E-learning is learning subjects with usage of internet technology. In the current proposed work E-learning websites were taken for providing suggestions to the user. The beauty of e-learning systems lie in the fact that they focus on learner's approach rather than the instructors approach. On the other hand e-learning systems suffer from a serious disadvantage that they offer the same learning material to all the learners. Although personalization is in being used and implemented widely but in the context of e-Learning the content being delivered to the learners is not profile oriented.

Mohamed et. al. [3] proposes a recommendation strategy in an e-learning environment that utilizes web usage mining for extracting the learners profile and information retrieval techniques for content based profiling. Using these profile links are suggested to an active learner accessing the website. The suggestions are provided to the user on the basis of collaborative filtering or content filtering or both.

Web mining and semantic web have been used for achieving personalization in e-learning by Sandesh et. al.[4]. The author categorizes audiences into four types namely individual learners, one who is focusing on a specific query, pragmatic and innovative learners. The issues pertaining to semantic web and web mining have also been discussed.

E-learning refers to the usage of electronic technology for teaching and learning purpose. It is termed as the most flexible, time saving and cheap mode of obtaining education. People are resorting to the needs of ICT for grabbing new technologies. Only a network platform and PC are required by learners. Some of the learner modules available are static and hence do not look upon specific learners queries. Web usage mining has been used by to personalize the web content thereby improving website design, customer satisfaction and user navigation. Association rule and sequential pattern mining have been utilized. Prefix span algorithm has been used in the current study and is proved efficient[5].

In another effort on Web usage mining in e-Learning Anuradha et. al. [6]proposed a model for enhanced e learning. The learners raise a mining request and sets mining parameters. After carrying out mining the results are provided to the user at his GUI.

Authors Richa et. al. in[7] focuses on designing of content for an e-learning site as content designing is considered very crucial these days. Content must be selected very nicely and must be updated on regular basis. The authors use the opinion of social clusters like academicians, researchers, students, alumni, industry for designing content. With the help

of a case study the results were shown. AHP or analytical hierarchy process, a technique of Multi criteria decision making was used for problem resolution. Later clustering was applied.

Ahmed et.al.in [8]in order to achieve personalization in e-learning , learning styles were integrated and an expert system has been proposed. Learning styles can help the students in achieving the desired results in much effective way. Mining of data can be accomplished by using several techniques. One such technique called Sequential pattern mining extracts patterns where support count is greater than or equal to minimum support threshold value. Authors Kapil et.al. [9] has elaborated on various data mining techniques with special emphasis on sequential pattern mining. Authors Irfan et.al. [10] have reviewed sequential pattern mining in detail. An overview of the features and techniques covered under Apriori and Growth pattern based have been given.

In another work on sequential pattern mining a sequence tree algorithm has been proposed by the authors [11]. The proposed algorithm works by finding out sequences that exist in the log file. Regular expressions can prove to be real powerful and popular tool for manipulating string data across various types of applications. Reg ExpMiner has been proposed which links together the RE and sequence mining worlds. The characteristics and expressiveness of the RE language, repetition of some characters in strings has been exploited while designing the above stated algorithm.

Regular expressions can be considered as a constraint for sequential pattern mining. Sequential Pattern Mining with Regular Expression Constraints (SPIRIT) algorithm has been proposed by Minos et.al. in [12] which mines frequent sequential patterns which also satisfy regular expressions constraints stated by user. Experimental study on synthetic and real life data sets has been carried out for showing the effectiveness of the proposed approach.

EXPERIMENTAL SETTING

An e-learning website has been designed which uses sequential pattern mining algorithm using regular expressions. The proposed website integrates content data and usage data and provide active recommendations to users. In order to judge whether the proposed work could satisfy the needs of the user in an e-learning website was built. A snapshot of the website is given below:

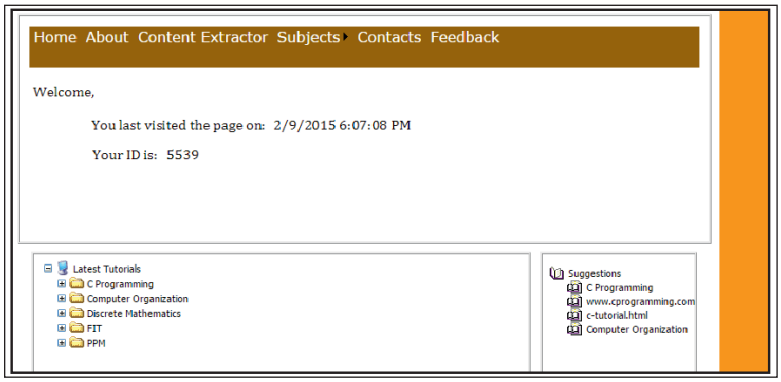
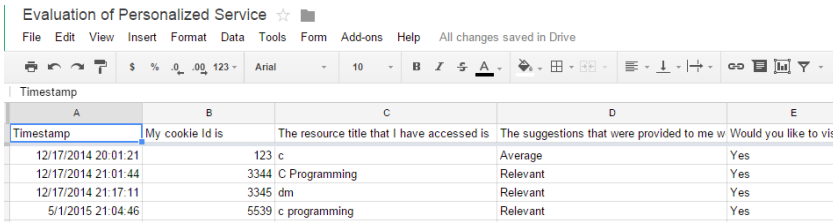


Fig. 3.0 Snapshot of Website

A group of 100 users was identified. The users identified were studying M.C.A. course and were in First Semester in GGSIPU since the website was primarily designed for only MCA First semester students. The students were asked to go through the website and examine its contents. They were requested to fill in the feedback form. The feedback form was posted on the website itself with an easy to use interface.

The feedback forms were later downloaded as Excel spreadsheets and were analyzed by applying data mining utilities. In the current study clustering, a data mining technique was utilized for analyzing the data. As stated earlier preprocessing of data was also carried out before applying data mining algorithms as mining can be applied only on preprocessed data. Mining has been carried out by using Data Mining Add ins available in Excel 2007 which performs mining in correspondence with SQL Server. A snapshot of the feedback form filled by the user visiting the website has been given below in Figure:

Figure 3.1 Snapshot of filled feedback form



Evaluation of Personalized Service

File Edit View Insert Format Data Tools Form Add-ons Help All changes saved in Drive

Timestamp

	A	B	C	D	E
	Timestamp	My cookie Id is	The resource title that I have accessed is	The suggestions that were provided to me w	Would you like to visi
	12/17/2014 20:01:21	123	c	Average	Yes
	12/17/2014 21:01:44	3344	C Programming	Relevant	Yes
	12/17/2014 21:17:11	3345	dm	Relevant	Yes
	5/1/2015 21:04:46	5539	c programming	Relevant	Yes

Figure 3.2 Snapshot of feedback form as recorded in Google Docs.

As shown above in the Figure the entries recorded are the cookie Id of the user, resource title accessed, whether the suggestion that were provided were termed satisfactorily by the users and finally whether one is interested in visiting the website again.

RESULTS AND DISCUSSION

After the responses were recorded from various users, these excel sheets were downloaded and clustering was applied. Cluster Analysis is a data mining technique which maximizes intra class similarity and minimizes inter class similarity. It is very popular and is available in almost all the data mining tools. Here clustering has been carried out using MS Excel Add-in for Data Mining. Clustering (in MS Excel Add-in for Data Mining) talks about the following:

Cluster Diagram: It is used for showing all the clusters present in a mining model.

Cluster Profiles: It gives an overview of all the clusters that are created by the algorithm. It displays all the attributes contained together with the distribution of attribute in every cluster. Histogram present shows the number of bars which can be increased or decreased depending upon the requirement. In the current work histogram bars option was set to 4. Generally the most important bars are shown.

Cluster Characteristics: This tab tells us about the characteristics which lead to the formation of the cluster. It comprises of three columns Variables, Value and Probability. Variables comprise of the attribute name, Values contain the possible value for that attribute and Probability contains the probability of that particular value.

Cluster Discrimination: In order to discriminate between any two clusters this tab is present. It helps in determining the most important differences between the two cluster along with the attribute variables and values which lead to this discrimination.

Cluster diagrams obtained after application of this technique on the downloaded

data have been given and the interpretations obtained have been discussed below:

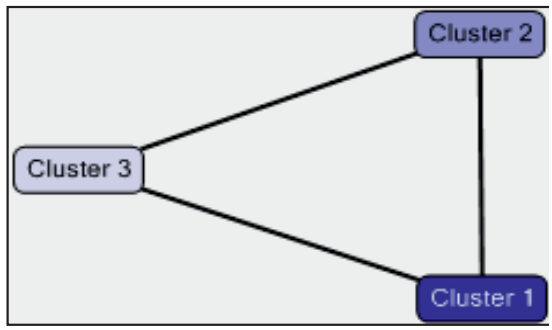


Figure 4.0 Cluster diagram for suggestions provided column

Drill through for model 'Range - Clustering_6'	
Cases Classified to: Cluster 3	
The suggestions that were provided to me were	**__RowIndex**
Irrelevant	7
Irrelevant	9
Irrelevant	14
Irrelevant	36
Irrelevant	48
Irrelevant	63
Irrelevant	66
Irrelevant	75
Irrelevant	100

Figure 4.1 Drill through model for cluster 3

Cluster Profiles

Variables	States	Population (All)	Cluster 1	Cluster 2	Cluster 3
Size		70	44	19	7
The suggestions that were provided to me were	Relevant	44	100 %	0 %	0 %
The suggestions that were provided to me were	Average	19	0 %	100 %	0 %
The suggestions that were provided to me were	Irrelevant	7	0 %	0 %	100 %

Figure 4.2 Cluster profile for suggestions column

Cluster Characteristics

Population (All)

Variables	Values	Probability
The suggestions that were provided to me were	Relevant	63 %
The suggestions that were provided to me were	Average	27 %
The suggestions that were provided to me were	Irrelevant	10 %

Cluster Characteristics

Cluster 1

Variables	Values	Probability
The suggestions that were provided to me were	Relevant	100 %

Figure 4.3 Cluster Characteristics

Cluster Characteristics
Population (All)

Variables	Values	Probability
The suggestions that were provided to me were	Relevant	63 %
The suggestions that were provided to me were	Average	27 %
The suggestions that were provided to me were	Irrelevant	10 %

Figure 4.4 Cluster Discrimination for suggestions column

As shown in Fig. 4.0 the values provided in suggestions retrieved have been divided into three clusters namely Cluster 1 (containing value relevant) , Cluster 2 (containing value Average) and Cluster 3 (with value Irrelevant). The shaded line present between the three clusters shows the strength of the similarity of clusters. Since a darker line is present in between all the three clusters, it represents a stronger association between the values contained in all the three clusters. The Cluster profiles given in Fig.4.2 clearly show the different values as 100% for Cluster 1, 2 and 3. The drill through model for Cluster 3 as shown in Fig. 4.1 shows that all the values rated Irrelevant have been placed in this cluster. It also represents that at which row the given value has been assigned. The next table i.e. Fig. 4.3 shows the characteristics of the various clusters. It shows the % probability assigned to every value if checked for the entire population. However, if it is checked for just cluster 1 the probability reflected is 100% which implies that Cluster 1 contains all the value rated as Relevant as shown in Fig.4.3. The next Fig.4.4 shows the Cluster Discrimination. It shows that Relevant Suggestions have been shown in

cluster 1 whereas the rest of the values have been shown in the complement of 1 cluster. Cluster diagram for visiting the website again

Figure 4.5 Cluster diagram

Cluster Profiles

Variables	States	Population (All)	Cluster 1	Cluster 2
Size		70	49	21
Would you like to visit the website again	Yes	49	100 %	0 %
Would you like to visit the website again	No	21	0 %	100 %

Figure4.6 Cluster profiles stating the % of values contained in each cluster

Cluster Characteristics

Population (All)

Variables	Values	Probability
Would you like to visit the website again	Yes	70 %
Would you like to visit the website again	No	30 %

Figure 4.7 Cluster characteristics showing the probability of each value

Cluster Discrimination

Variables	Values	Favors Cluster 1	Favors Complement of Cluster 1
Would you like to visit the website again	Yes		
Would you like to visit the website again	No		

Figure 4.8 Cluster discrimination for each cluster

Once the visitor has visited the website he decides on his mind whether the website would be visited or not. This is directly dependent on the satisfactory responses that one gets from the website. In the feedback it was also included whether the visitor would visit the site or not. The responses were generated in the form of Yes or No. The cluster diagram presented in Fig. 4.5 shows the two clusters which are formed for the two values. A darker

value in cluster indicates a higher inclination for the values contained in cluster 1. The cluster profiles shown in Fig. 4.6 show clearly that cluster 1 contains 'YES' response whereas cluster 2 contains 'NO' as response. A 100% probability has also been shown against Cluster 1 for 'YES' and in cluster 2 for 'NO'. Fig. 4.7 shows the cluster characteristics. Here it has been shown clearly that the probability of 'YES' is 70% whereas for 'NO' it is 30%. Fig. 4.8 i.e. the cluster discrimination tells us that the value YES is contained in cluster 1 whereas, the other value NO is contained in the complement of cluster 1.

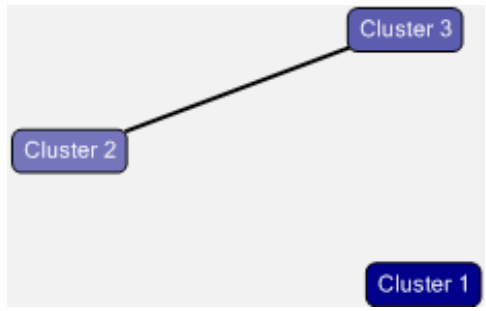


Figure 4.9 Cluster diagram for resource title accessed column

Cluster Profiles

Variables	States	Population (All)	Cluster 1	Cluster 3	Cluster 2
Size		70	31	21	18
The resource title that I have accessed is	C Programming	31	89 %	0 %	0 %
The resource title that I have accessed is	FIT	14	2 %	13 %	62 %
The resource title that I have accessed is	PPM	11	0 %	59 %	5 %
The resource title that I have accessed is	Computer Organisation	10	6 %	26 %	20 %
The resource title that I have accessed is	Discrete Mathematics	4	4 %	2 %	13 %

Figure 4.10 Cluster profiles indicating the % of each value contained in respective cluster

Cluster Characteristics
Population (All)

Variables	Values	Probability
The resource title that I have accessed is	C Programming	44 %
The resource title that I have accessed is	FIT	20 %
The resource title that I have accessed is	PPM	16 %
The resource title that I have accessed is	Computer Organisation	14 %
The resource title that I have accessed is	Discrete Mathematics	6 %

Figure 4.11 Cluster characteristics highlighting the probability of each value

Cluster Discrimination

Variables	Values	Favors Cluster 1	Favors Complement of Cluster 1
The resource title that I have accessed is	C Programming		
The resource title that I have accessed is	FIT		
The resource title that I have accessed is	PPM		
The resource title that I have accessed is	Computer Organisation		

Figure 4.12 Cluster discrimination showing which values favor cluster 1

The cluster diagram given in Fig. 4.9 shows that the values contained in cluster 1 are completely isolated and have no relation with values in Cluster 2 or 3. A shaded line represents the strength of the similarity of the cluster. Since no shaded line is present from cluster 1 to cluster 2 or cluster 3 no association exists. The cluster profile present in Fig. 4.10 highlight that a majority of people have accessed the title 'C Programming' and hence have been placed in cluster 1. The other resources that have been accessed are placed in cluster 2 and 3 as shown in cluster profiles. The fig.4.11 i.e. cluster characteristics shows the probability of each resource title. It shows that the maximum accessed resource is 'C Programming' with a % of 44% whereas 'Discrete Mathematics' has been accessed the least with a probability of 6%. The cluster discrimination given in Fig. 4.12 shows that only the title 'C programming' is favored in cluster 1 rest all others are contained in either cluster 2 or 3.

CONCLUSION

In the current work Clustering; a technique of data mining has been applied on feedback of the users available from an e-learning website. The technique has been applied to check whether the website which has been developed is satisfying the requirements of the users or not. The inferences drawn have been discussed and can be used for further refinements in the site. It has been noticed that majority of the users were pleased with the suggestions provided and were interested in visiting the website again. However it was found that not all the resource materials provided in the website were found attractive. The resource title C Programming was the most preferred. On the other hand Discrete Mathematics was the least accessed resource.

As part of the future work we can incorporate the results obtained during our study in the website. The effectiveness of the website can be checked once again. Clustering has been observed to be more effective as it provides in depth analysis.

References

Neha Goel, C.K. Jha ,"Preprocessing Web logs: A Critical phase in Web Usage Mining " presented in 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, IEEE Explore, 2015, pp. 672-676.

Neha Goel, Sonia Gupta and C.K. Jha, "Analyzing Web Logs Of An Astrological Website Using Key Influencers" in IARS International Journal, Vol. 05 No. 01 2015, p-ISSN 2202-2821 e-ISSN 1839-6518.

Mohamed Koutheaïr Khribi, Mohamed Jemni and Olfa Nasraoui, "Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval " , *International Forum of Educational Technology & Society*, 2009 , pp.30-42.

Sandesh Jain, Dhanander K. Jain, Harihar Bhojak, Ankit Bhilwar, and Mamatha J.," Personalization of e-Learning Services using Web Mining and Semantic Web", *International Journal of Machine Learning and Computing*, Vol. 2, No. 5, October 2012,pp.569-572.

Xue Sun, Wei Zhao, " Design And Implementation Of An E-Learning Model Based On Wum Techniques", *International Conference on E-learning, E-business*, EIS and E-Government, IEEE, 2009,pp. 248-251.

Anuradha Yadav, Satbir Jain," Analyses Of Web Usage Mining Techniques To Enhance The Capabilities Of E-Learning Environment", *IEEE*, 2011, pp. 223-225.

Richa Sharma, Hema Banati and Punam Bedi, "Incorporating Social opinion in content selection for an e-learning course ", *ICCSE* 2011, IEEE, 2011,pp. 1027-1032.

Ahmed Al Hamad, Norlaily Yaacob and A.Y. Al-Zoubi," Integrating 'LearningStyle' Information into Personalized e-learning System, *IEEE Multidisciplinary Enginnering Education Magazine*, March 2008, pp.2-6.

Kapil Sharma, Ashok, Dr. Harish Rohil, "A Study of Sequential Pattern Mining Techniques *", International Journal of Engineering and Management Research*, Volume-4, Issue-1, February-2014, ISSN No.: 2250-0758, Page Number: 241-248.

Irfan Khan, Anoop Jain, "A Comprehensive Survey on Sequential Pattern Mining ",*International Journal of Engineering Research & Technology* (IJERT) Vol. 1 Issue 4,

June – 2012, ISSN: 2278-0181, pp.1-6.

Priyanka Baraskar, Supriya Chavan, Dipshree Dhage, Samruddhi Giri, Jayashree Jha, "Extracting Frequent Sequences from Web Log Data using Sequence Tree Algorithm ",*International Journal of Advanced Research in Computer and Communication Engineering* Vol. 4, Issue 3, March 2015,pp. 10-13.

Minos N. Garofalakis, Rajeev Rastogi and Kyuseok Shim, "SPIRIT: Sequential Pattern Mining with Regular Expression Constraints ", *Proceedings of the 25th VLDB Conference*, Edinburgh, Scotland, 1999, pp. 223-234.

G.K. Gupta, "Web Data Mining," in *Introduction to Data Mining with Case Studies*, India: PHI, 2011, ch.5, sec.5.1-5.5, pp. 231-250.

Lingma Lu Acheson and Xia Ning , in" Enhance E-Learning through Data Mining for Personalized Intervention" in *Proceedings of the 10th International Conference on Computer Supported Education* (CSEDU 2018) - Volume 1, pages 461-465.

Margaret H. Dunham, "Web mining," in Data Mining Introductory and Advanced Topics, *India: Pearson Education*, 2006, ch.7, sec. 7.2, pp. 194- 202.