EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

Features Analysis of Online Shopping System Using WCM

Maria Erum[1], Muhammad Waqas[1], Sidra Arshad[1], and Tahir Nawaz[1]

[1]Department of Computer Science, The University of Lahore, (Sargodha Campus)

tahir.nawaz@cs.uol.edu.pk

Abstract

Data mining techniques being used for web information extraction are unbelievable systems and suggested for the protection of extremely susceptible data. By the web sources huge amount of data is maintained and can be easily retrieved by using the web mining techniques as the techniques are applied exactly based on the needs of the users. ECommerce and online shopping noticed a huge growth in business industry. This facility has been mostly employed in western countries during the last two decades. In east online shopping is increasing as most of the business is running through web site as well as in west. This business can be boost by feature analysis of different successful running online web stores. This study is going to present analysis of different features of successful online business website and of those which are not that much popular and accessed infrequently, their features will be extracted and compared to get the reasons of popularity of frequently accessed online shopping websites and after that recommendations will be made to increase the traffic of unpopular online shopping websites to dominate online business in Pakistan. According to the presented work it has been concluded that unpopular websites lack some features that are included in popular websites such as brands, as people are more conscious about brands and labels so they visit and shop from the websites which offer them best quality famous brands, moreover it has been observed that unpopular websites have less categories they must broaden their variety of products especially related to sports, fitness, bathroom accessories, technology and cosmetics. Apart from these another interesting fact that has been found is that popular websites mostly attract their customers by giving them offers such as buy one get one free, free home delivery and free gifts, such offers are always attracting new and more customers. Unpopular websites can improve their business by including the features discussed above. The results of research on this dataset also show that the Naïve bayes is better than j48 in terms of efficiency and accuracy respectively.

Keywords: WCM (Web Content Mining), Online Shopping, Classification, Data Mining Techniques, Decision Tree, Naïve Bayes, J48.

Received on 19 November 2017, accepted on 12 February 2018, published on 13 April 2018

doi: 10.4108/eai.13-4-2018.154471

1. Introduction

WWW is a collection of massive data. The web is very enormous, diverse, flexible and dynamic. Continuous expansion of web with respect to amount of traffic and size and complexity of websites, it's becoming a great obscurity to find the suitable and relevant information from the web. Content of web mostly consist of unstructured and heterogeneous information which is hidden in web. Web mining is a promising field that aims to find and extract the relevant and valuable information that is hidden in the data related to the web. Data mining is the process of finding and extraction of meaningful and valuable information from large amount of data. Today every type of data is on the web sites, such as text, video, images, hyperlinks, audio, and metadata. On the basis of such diversity the web mining further divided into three categories which are Web Usage mining, WCM, and Web structure mining figure 1.1.

Web usage mining deals with the web logs and search histories of different users while they interact with web.

Web usage mining involves user access patterns from one or more web servers for the automatic discovery. This analysis is used for classification such as site amendments, personalization, system enhancements, business intelligence and usage classification [1].
WCM targets to mine the content of the web such as text, images, audio, video and results of search. Further classification is search result mining and web page content mining [3].
Web structure mining aims to deal with hyperlinks with the web itself. Structure of most web graphs consists of hyperlinks as edges and web pages as nodes, hyperlinks behave like edges between two related web pages [2].
Web mining can be further divided into following sub tasks
1) Resource finding: in this task the required web documents are retrieved.
2) Information selection and pre-processing: automatic selection and preprocessing of specific information from extracted web documents
3) Analysis: interpretation and/or validation of patterns mined from web [4].

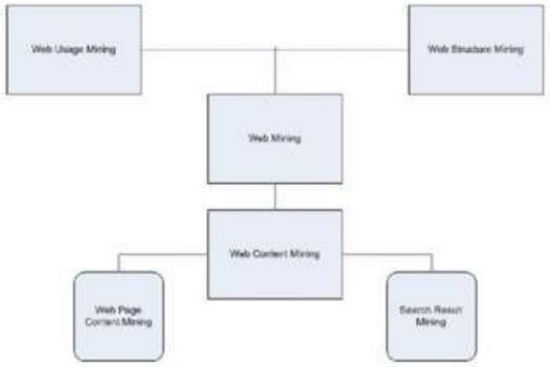

Fig2.1 Classification of Web Mining [3]

Web mining is a technique to discover unknown and undiscovered patterns of users of web. These patterns give us a lot of information which is later transformed in to the knowledge and we use this knowledge increase our business [5]. Learning about users patterns gives us an overview of user behaviors on the web, which may lead us to web personalized as our user's usage. Web mining a sub category to find unknown knowledge from WWW [2]. Many other techniques were applied in order to retrieve information and extracting information form huge data resides on the web, comparison of those techniques are given in [5]. Selection of useful information is done after indexing the text [6].Extractions of information depend upon selected relevant data and facts whereas selecting relevant document is done by information retrieval. Web mining have become part of Information Extraction System (IES) and Information Retrieval System (IRS). IES is pre-processing stage before mining is applied to data which also index text to retrieve data. Machine Learning (ML) is not a part or directly has connection with web mining but it help to improve and gets better text classification than retrieval system [7].

1.1Web content Mining

Traditionally content of web used to search by content of web. WCM tends to be work like search engine. It is the process of extraction of relevant knowledge available on the web in form of data. This mining concerned with extraction of relevant text by removing noise like navigational elements, advertisements, copy right notes and contacts. Growing applications of WCM includes Automatic extraction of semantic relations and structure from web.

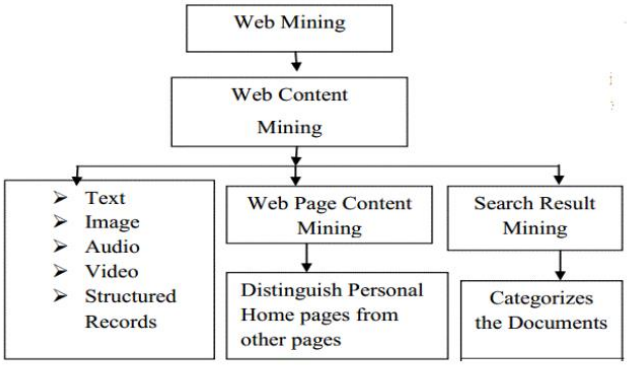

Fig 2.2WCM Taxonomy [8]

Two approaches are being used in WCM, first one is Agent based and second one is database approach. Agents based approach further divided into three kinds of agent's names as, personalized web agents, categorizing/filtering agent, and intelligent agents [3]. Mining of multimedia, semi structured, structured, and unstructured data makes the WCM a complicated task. Web content mining sub categories are shown in figure 1.2.

2. Literature Review

Online shopping systems information extraction assist to find hidden information from the vast amount of products such as the product features and its specification. In earlier days the techniques used for the information extraction from web documents were based on the HTML documents. Based on that HTML document a tree is formed of a web page. Information is retrieved through the search methodologies of a tree. The leaf node must be a text node which is to be mined from the product. Extraction is performed by parsing through Hidden Markov Model and then it classifies the information needed. This model was used to learn the attributes automatically.
Gengxin Miao (2009) focuses on the list of objects that appears repeatedly based on the tag paths in the DOM tree of the respective web documents. Then based on the comparison of the occurrence patterns of the tag paths the visually appearing signals are identified and clustering is performed based on the similarity measures of tag paths. This method had higher accuracy when comparing to previous methods [10].

Wei Liu (2010) presents an approach that extracts the products and its specifications from the online shopping web sites based on the visual features. All the visual features are considered such as content feature, format feature etc of the text document and clustered ba sed on the similarity measures. This implementation also takes the DOM tree for data records extraction. From that extracted record the data items which are the product information can be retrieved [11].

Ali Ghobadi (2011) presents an improved web information extraction which is based on ontology. To extract the attributes that is of semantic meaning the ontology method of label identification for attributes are used. These processes make use of assumptions on information and fully understand the semantics of the HTML documents and extract the information automatically [12].

Xiaoqing Zheng (2012) introduced structural semantic entropy used for locating the data of interest in a web page, based on the measurement of the density of occurrence of the relevant information. Due to the difficulty of writing and, maintaining the wrappers and blocks identification in the vision based extractors this method has been introduced. Entropy measure is calculated to identify the density of the product specified and labeled [13].

3. Methodology

The methodology used for this study is shown in Fig 4.1. The major steps are data collection, preprocessing (i.e., attribute selection, feature weighting and tokenization), feature selection, classification and evaluation. These steps are described below.

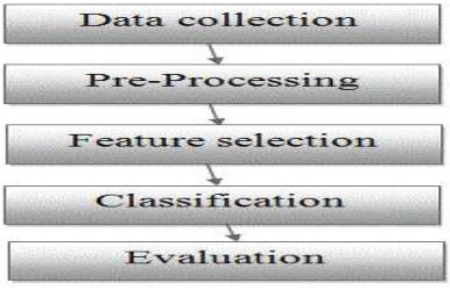

Fig 4.1 Steps of Research methodology

3.1 Data Collection

Our data set consisted of text extracted from popular and unpopular online shopping websites. Text files consist of data from home pages and further 9 more categories mentioned. In order to create the dataset, services of web crawler, popular and unpopular online shopping websites were used.

3.2 Popularity comparison between online shopping websites

The comparison between the popularity among different online shopping websites has been carried out by online comparison of number of unique visitors on daily bases using web traffic analytic.

3.3. Data Pre-Processing

Before feature classification method some pre-processing steps are necessary. These steps consist of tokenization; feature weighting and removal of stop words. There are number of tokenization techniques available such as, phrase level, word level, and sentence level.

3.4. Feature selection

First and most important process is pre-processing in classification and pattern recognition in data mining are Feature extraction or selection. It's considered as efficient preprocessing technique for eliminating noise and it also reduces dimensionality.

Feature selection is major significant step in feature classification. In feature selection we struggle to get rid of worthless words from the text to increase classification correctness and to reduce computational hurdles.

3.5. Classification Methods

The classification algorithm learns from the training set and builds a model. The built model is employed to classify new items. We have posed feature analysis of online shopping websites as feature classification assignment where text based dataset is applied to classifiers as input then tokenized the dataset structure and output is the counts of each word or token. Whole idea is presented in the diagram Fig 3.2. We have selected two algorithms J48 a tree based algorithm and Naive Bayes (NB). A lot of empirical studies have been carried out to comparatively evaluate the efficiency of the algorithms.

4. Confusion matrix:

Confusion matrix includes information about predicted and actual classification. Confusion matrix depicts the accurateness of the result to a classification problem. Given n classes a confusion matrix is a m x n matrix, where Ci,j indicates the number of tuples from D that were assign to class Ci,jbut where the correct class is Ci.Obviously the best solution will have only zero values outside the diagonal. Performance of such systems is normally assessed using the data given in the matrix. The

entries in the confusion matrix have the following Table5.1 meaning in the context of our study:

Table 5.1 Functions and their details

1. Functions	1. $Precision = \frac{TP_i}{TP_i+FP_i}$ Formula	1. Description
1. Precision 2.	2.	2. In mining precision is called positive predictive value of relevant instances. 3.
1. Recall 2.	3. $Recall = \frac{TP_i}{TP_i+FN_i}$	4. Recall is the proportion of retrieved relevant instances in fraction. 5.
1. F-Measure 2.	4. F-measure = $\frac{2\ PR}{P+R}$ 5.	6. A collective measure of recall and precision is calculated as Recall * Precision * 2/ (Recall + Precision). 7.

a. experimentation Tool: WEKA

WEKA (acronym of "Waikato Environment for Knowledge Analysis").We have chosen WEKA software for our experimentations. As it includes all the essential functionalities for this work. For example, it includes feature selection method IG, stop words removal; attribute selection and classification method etc.

5. Experimentation and Results

The results shown in this section is purely in terms of the classifier success rate over given dataset. The experimentation flow can be seen in the Fig 6.1. In this study we have used two algorithms for our classifying tasks. For characterization of classifiers 10 cross standard validation is used. It is a technique to generalize independent set in statistical results.

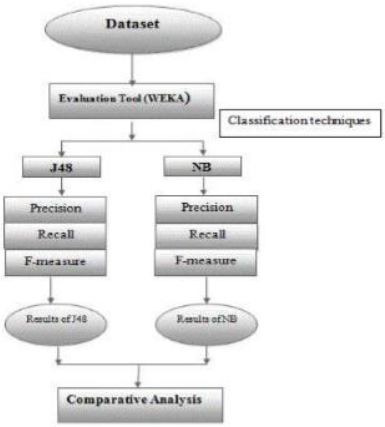

Fig6.1 Experimental Model

6.3. Results

Table6.1 Confusion Matrix of J48:

1. A	1. B	1. <-- classified as
1. 85	2. 8	2. a = pop
1. 12	3. 35	3. b = unpop

For above confusion matrix, true positives for class a='POP' is 85 while false positives is 8 whereas, for class b='UNPOP', true positives is 35 and false positives is 12 i.e. diagonal elements of matrix 85+35 =120corresponds to the instances classified correctly and other elements 12+8 = 20 represents the incorrect instances. True positive rate = diagonal element/ sum of relevant row False positive rate = non-diagonal element/ sum of relevant row Hence,

Table 6.2 Effectiveness of J48

1. TP rate for class a	1. 85/(85+8) = 0.914
1. FP rate for class a	2. 12/(12+35) = 0.255
1. TP rate for class b	3. 35/(12+35) = 0.745
1. FP rate for class b	4. 8/(85+8) = 0.086
1. Average TP rate	5. 0.857
1. Average FP rate	6. 0.198

1. Precision for class a	7. 85/(85+12) = 0.88
1. Precision for class b	8. 35/(35+8) = 0.81
1. F-measure for class a	9. 2*0.88*0.914 /(0.88+ 0.914)= 0.893
1. F-measure for class b	10.2*0.8* 0.745 /(0.8+ 0.745) = 0.778
1. Correctly Classified Instances	11. 120 85.7143 %
1. Incorrectly Classified Instances	12. 20 14.2857 %

Table 6.2 depicts the different measures obtained by confusion matrix of j48

Table 6.3 Confusion Matrix of Naïve Bayes

1. A	1. B	1. <-- classified as
1. 80	2. 13	2. a = pop
1. 6	3. 131	3. b = unpop

For above confusion matrix, true positives for class a='POP' is 80 while false positives is 13 whereas, for class b='UNPOP', true positives is 131 and false positives is 6 i.e. diagonal elements of matrix 80+131 =211 corresponds to the instances classified correctly and other elements 6+13 = 19 represents the incorrect instances. True positive rate = diagonal element/ sum of relevant row False positive rate = non-diagonal element/ sum of relevant row Hence,

positive rate = non-diagonal element/ sum of relevant row Hence,

Table 6.4 Effectiveness of Naïve Bayes

1. TP rate for class a	1. 80/(80+13)=0.860
1. FP rate for class a	2. 6/(6+131) =0.128
1. TP rate for class b	3. 0.872
1. FP rate for class b	4. 0.14
1. Average TP rate	5. 0.864
1. Average FP rate	6. 0.132
1. Precision for class a	7. 80/(80+6) = 0.930
1. Precision for class b	8. 131/(131+13) = 0.759
1. F-measure for class a	9. 2*0.93*0.86/(0.93+0.86)= 0.896
1. F-measure for class b	10.2*0.759*0.872/(0.759+0.872) = 0.811
1. Correctly Classified Instances	11. 121 86.4286 %
1. Incorrectly Classified Instances	12. 19 13.5714 %

Classification results using NB and j48 algorithms are given below in table 6.5.

Table 6.5 Comparison of J48 and Naïve Bayes

1. Evaluation criteria	1. J48	1. Naïve Bayes
1. Time to build model in second	2. 2.06	2. 0.64
1. Correctly classified instances	3. 120	3. 121
1. Incorrectly classified instances	4. 20	4. 19
1. Prediction accuracy	5. 0.855	5. 0.866

1. Cost	6. Unpop	6. 0.867	1. 0.948
	7. Pop	7. 0.867	2. 0.943

Table 6.5 shows the summary of comparison between j48 and Naïve Bayes, which clearly depicting that the Naïve Bayes performed much better than the J48 in terms of accuracy, prediction and time to build model.

Graphs using Naïve Bayes

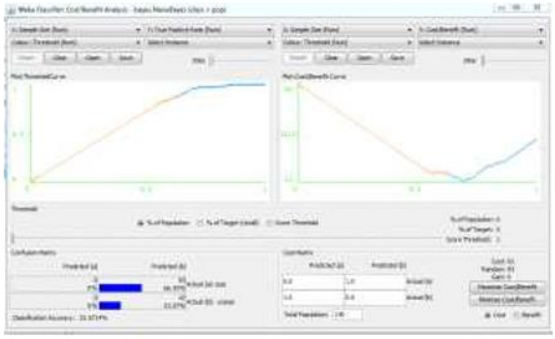

Fig 6.2 ROC Naïve Baayes (popular)

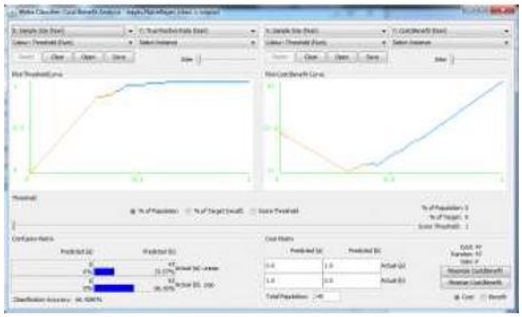

Fig6.3 ROC Naïve Baayes (Unpopular)

The above shown graphs represents Confusion matrix, True positive rate , False Positive Rate, Precision, Recall, ROC Area, F-Measure on the basis of class popular and unpopular. There are two graphs for each class .Graph 1st is about sample size and True positive Rate. Graph 2nd is for sample size and cost benefit analysis. These graphs are drawn on the basis of confusion matrix and cost matrix. In this we are using Naïve Bayes classification algorithm.

Graphs using j48

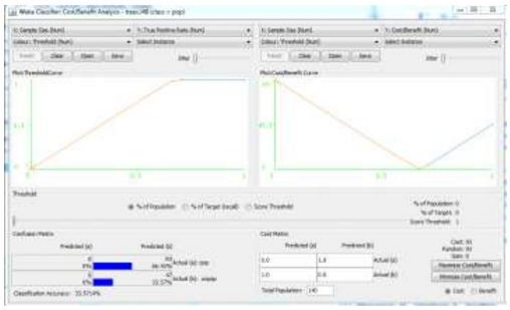

Fig 6.4 ROC (J.48 popular)

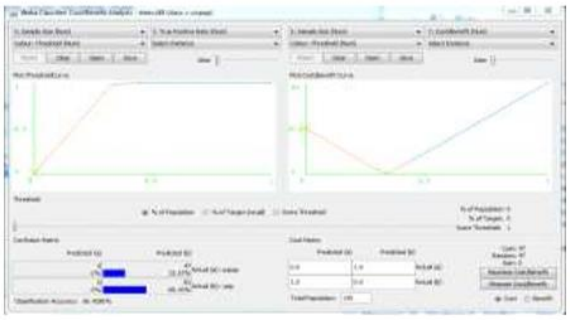

Fig 6.5 ROC (J.48 Unpopular)

The above shown graphs represents Confusion matrix, True positive rate , False Positive Rate, Precision, Recall, ROC Area, F-Measure on the basis of class popular and unpopular. There are two graphs for each class .Graph 1st is about sample size and True positive Rate. Graph 2nd is for sample size and cost benefit analysis. These graphs are drawn on the basis of confusion matrix and cost matrix. In this we are using J48 classification algorithm.

A comparison of classifiers for feature classification is given in Table 6.6. For feature classification NB have provided approximately 90 % F - measure and 87% precision which is best as compared to results of J48.

Table 6.6 Comparison of Results of Classifications

1. Classifiers	1. Precision	1. Recall	1. F-measure
1. NB	2. 0.87	2. 0.86	2. 0. 896
1. J48	3. 0.85	3. 0.85	3. 0.893

In the light of above results shown in Table 6.6, we can conclude that NB performed best for feature classification. This also shows that NB has edge over other classifiers and suitable for designing Feature Analysis and Recommendation system.

Table 6.7 Availability of Words Popular VS Unpopular by NB

1. SR	1. Word	1. Popular	1. Unpopular
1. 1	2. Cricket	2. 1	2. 0
1. 2	3. Gucci	3. 1	3. 0
1. 3	4. Size	4. 1	4. 1
1. 4	5. Squash	5. 1	5. 0
1. 5	6. Wholesale	6. 1	6. 0
1. 6	7. Tennis	7. 1	7. 0
1. 7	8. Total	8. 1	8. 0
1. 8	9. Pc	9. 1	9. 0
1. 9	10.Calvin	10.1	10.0
1. 10	11.Descending	11.0	11.1
1. 11	12.Keyboards	12.0	12.1
1. 12	13.QuadCore	13.0	13.1
1. 13	14.FREE	14.1	14.0
1. 14	15.View	15.0	15.1
1. 15	16.Straighteners	16.0	16.1
1. 16	17.Juicers	17.1	17.0
1. 17	18.Mufflers	18.1	18.0
1. 18	19.Alkaram	19.1	19.0
1. 19	20.Combo	20.1	20.0
1. 20	21.Delivery	21.1	21.1

In Table 6.7, 20 Different words are displayed which shows the effects of these word among popular and unpopular websites. In above table value One '1' means that this word is available and value Zero '0' is missing in the respective category. As shown we have selected only twenty words for demonstration from more than 10000 words. We selected words like, Calvin, Alkaram and Gucci are the brands name which are present popular websites but are missing in unpopular websites

Similarly popular websites are providing more products than unpopular websites such sports facilities (Tennis, Squash, and Cricket) which are not present in unpopular. The word 'Free'in the table is presented in popular category but not in the unpopular, this could be free delivery or free gift for customers on purchase which attracts more users to their site.

Table 6.8 Words StdDev

1. Sr	1. Word	1. Popular	1. Unpopular
1. 1	2. Cricket	2. 0.5	2. 0.4107
1. 2	3. Squash	3. 0.4707	3. 0.444
1. 3	4. Tennis	4. 0.5	4. 0.4107
1. 4	5. Pc	5. 0.2838	5. 0.7425
1. 5	6. Calvin	6. 0.5178	6. 0.1901
1. 6	7. FREE	7. 0.3552	7. 0.1215
1. 7	8. View	8. 0.3639	8. 0.7286
1. 8	9. Delivery	9. 0.4218	9. 0.4656
1. 9	10.Alkaram	10.0.5694	10.0.258
1. 10	11.Combo	11.0.5707	11.0.266

In table 6.8 Standard Deviation of the words are given whose has an effect on popular websites in getting more hits than other unpopular websites.

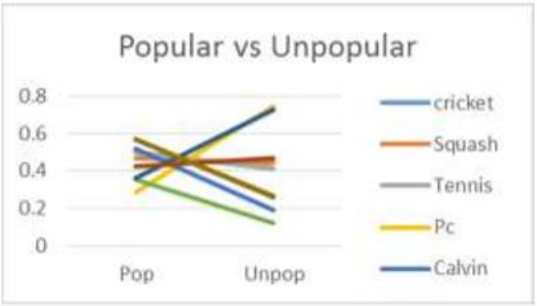

Fig 6.6 Popular website VS. Unpopular websites

In fig 6.5 a line chart shows the change of values in graphical form of unpopular and popular websites.

Conclusion

In this research a Features Analysis of Online Shopping System Using WCM has been presented. The results have presented with the help of two popular classifiers algorithm and compared their performance over given dataset.10 folded cross validation has been used in order to validate the results. Results show that our approach for feature analysis is very effective. This study will help the researchers to understand the trend of e-commerce and encourage them to work on content mining to get more depth knowledge of ongoing business in order to get better results in future.

According to the results of this study it is been concluded that unpopular websites lack brands, as people are more conscious about brands and labels so they visit and shop from the websites which offer them best quality famous brands, moreover it has been observed that unpopular websites have less categories they must broaden their variety of products especially related to sports, fitness, bathroom accessories, technology and cosmetics. Apart from these another interesting fact that has been found is that popular websites mostly attract their customers by giving them offers such as buy one get one free, free home delivery and free gifts, such offers are always attracting new and more customers. Unpopular websites can improve their business by including the features discussed above.

The future work includes creating of self-ruling specialists that break down the found standards to give important approaches or proposals to clients. Future extent of WCM incorporates anticipating client needs with a specific end goal to enhance the ease of use, adaptability, client maintenance, and confining a productive structure for Web Personalization through productive utilize Web Log Files.

Future Directions

We have selected shopping sites running in Pakistan. Be that as it may, this review can be further prompt to various substance of different sorts of sites, for example, stimulation and facilitating destinations. Doing this, we can make a suggestion framework for setting parameters to get fame among online business regarding site content. It will help the web designers and agents to create online organizations in an approach to get more mainstream among clients. Furthermore different looks into can without much of stretch discover the dataset of mainstream and disagreeable Online Shopping sites of Pakistan which data can be gathered efficiently. As trained once, this classified model can be further use for prediction.

References

[1] M.chaturvedi, J.Vidyapeeth, A SURVEY ON WEB MINNING Algorithms, Sandhya (M.tech CSE), Jaipur, solan, 25 March 2013.

[2] C.Menaka, N.Nagadeepa, A Survey of WCM Tools and Future Aspects, 13-Aug-2014, ISSN_NO: 2321-3337.

[3] F. Johnson, S.K. Gupta, WCM Techniques: A Survey, Volume 47– No.11, June 2012.

[4] R. Sarla*et al.,* Analysis & Approaches of Web Mining Prof. Rajesh Shah Research Scholar, December 2015

[5] B. Liu, K. C. Chiang, Editorial Special Issue onWCM,Acm. Journal of Machine Learning Research 4, 177-210.

[6] W. Liu *et al.,* ,"ViDE: A Vision -based Approach for Deep Web Data Extraction", IEEE Transactions on Knowledge and Data Engineering, Volume:22 , Issue: 3, Mar ch 2010,pp.447 – 460.

[7] J.Guo, V.Keselj and Q.Gao, Integrating Web Content Clustering Into Web Log Association Rule Mining,SpringerVerlag, Vol. 3501 Lnai, 182-193, 2005.

[8] G. Miao *et al.,* " Extracting Data Records from the Web Using Tag Path Clustering",International World Wide Web conference Committee (IW3C2),April,2009,pp.981- 990.

[9] V.Gedov*et al.,* Matching Web Site Structure andContent,Acm. Proceedings ofthe 13th International World Wide Web Conference on Alternate Track Papers and Posters, 2004.

[10] Gengxin Miao,Junichi Tatemura, Wang -pin Hsiung, Arsany Sawires, Louise E.Moser,"Extracting Data Records from the Web Using Tag Path Clustering",Interna tional World Wide Web conference Committee (IW3C2),April,2009,pp.981- 990.

[11] Wei Liu, Xiaofeng Meng , Weiyi Meng ,"ViDE: A Vision -based Approach for Deep Web Data Extraction", IEEE Transactions on Knowledge and Data Engineering, Volume:22 , Issue: 3, Mar ch 2010,pp. 447–460.