Web page Classification Techniques-A Comprehensive Survey

[1]Sonal D. Vaghela, [2]Pinal Patel
[1]PG Student, [2]Assistant Professor
Department of Computer Science and Engineering, Government Engineering College, Gandhinagar, India
[1]sonalvaghela@gmail.com , [2]pinalpatel@gecg28.ac.in

***Abstract*— Today, the world is moving across the internet. Internet is collection of Web Pages. Web Page contains a bunch of information. In bunch of information to find or retrieve particular page or information is difficult task. It is difficult for the Search Engine to identify web page. So make this task easy there are different web page classification methods.web page classification is a web mining area. Using this method we can identify web pages. Web page Classification retrieves WebPages based on content and structure of web page. This paper shows results of different Classification methods and comparison of that**

***Index Terms*— web page classification, SVM, Naïve Bayes, Extraction, Feature**

I. INTRODUCTION

The fast developments on the computer and networking technologies have increased the popularity of the Web which has caused the inclusion of more and more information on the Web [10]. Web contents, including online documents, e-books, journal articles, technical reports and digital libraries, have been rapidly exploring all time. It is much helpful to categorize web contents for efficiently contents browsing, managing, even spam filtering [5].

Web page Classification is a process where one page is appended to one or more directories which is predefined in advance [3]. Automatic Web page classification is a supervised learning problem in which a set of labeled Web documents is used for training a classifier, and then the classifier is employed to assign one or more predefined category labels to future Web pages [2]. Most of the applied web page classification techniques are inherited from automatic text classification: a supervised learning task, defined as assigning pre-defined category labels to new documents, based on the likelihood suggested by a training set of labeled documents. Therefore, an increasing number of learning approaches have been applied to classify web pages [11].

Web page classification techniques use concepts from many fields like Information filtering and retrieval, Artificial Intelligence, Text mining, Machine learning techniques and so on. In the machine-learning model, a classifier is given training with already classified examples and it learns the rules for classification during this training phase. Then this classifier is used for classification of the new pages. A machine learning classifier shows improved performance with experience as it learns every time it classifies the page. There are many machine learning algorithms used for web page classification – decision lists, decision trees, artificial neural networks, Bayesian classifiers, SVM classifiers,k_nearest neighbor, clustering algorithms etc.web page classification problem as it supports the dynamic nature of the web pages[7].

II. WEB PAGE CLASSIFICATION APPLICATION

Applications of Web Page Classification are [12]:

a. Constructing, maintaining or expanding web directories (web hierarchies)
b. Improving quality of search results
c. Helping question answering systems
d. Building efficient focused crawlers or vertical search engines
e. Web content filtering
f. Assisted web browsing
g. Knowledge base construction

III. WEB CONTENT MINING

Web content mining targets the knowledge discovery, in which the main objects are the traditional collections of text documents and, more recently, also the collections of multimedia documents such as images, videos, audios, which are embedded in or linked to the Web pages as shown in figure 2. Web content mining could be differentiated from two points of view: the agent-based approach or the database approach. The first approach aim son improving the information finding and filtering and could be placed into the following three categories [14]:

i. Intelligent Search Agents: These agents search for relevant information using domain Characteristics and user profiles to organize and interpret the discovered information.
ii. Information Filtering/ Categorization: These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize those.
iii. Personalized Web Agents: These agents learn user preferences and discover Web information based on these preferences, and preferences of other users with similar interest. The second approach aims on modeling the data on the Web into more structured form in order to apply standard database querying mechanism and data mining applications to analyze it. The two main categories are multilevel databases and Web query systems. The figure 2 shows that the

overview of web content mining [13].

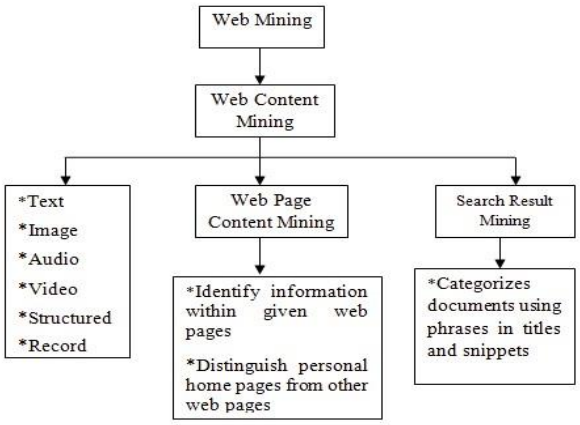

Figure 1: Web Content Mining

IV. WEB CONTENT MINING TECHNIQUES

The two common tasks through which useful information can be mined from Web are Clustering and Classification. Here present various classification algorithms used to fetch the information. Classification is often posed as a supervised learning problem in which set of labeled data is used to train a classifier which can be applied to any label future example[15].

A) Decision Tree

Decision tree is a powerful classification technique. The decision trees, take the instance described by its features as input, and outputs a decision ,denoting the class information in our case .Two widely known algorithms for building decision trees are Classification and Regression Trees and ID3/C4.5.The tree tries to infer as split of the training database on the values of the available features to produce a good generalization. The splitatach node is based on the feature that gives the maximum information gain. Each leaf node corresponds to a class label. A new example is classified by following a path from the root node to a leaf node, where a teach node attest is performed on some feature of that example. The leaf node reached is considered the class label for that example .The algorithm can naturally handle binary or multiclass classification problems. The leaf nodes can refer to either of the K classes concerned.

B) k-Nearest Neighbor

kNN is considered among the oldest non-parametric classification algorithms. To classify an unknown example the distance (using some distance measure e.g. Euclidean) from that example to every other training example is measured. The k smallest distances are identified, and the most represented class in the se k classes is considered the output class label. The value of k is normally determined using a validation set or using cross-validation.

C) Naive Bayes

Naïve Bayes is a successful classifier based upon the principle of Maximum A Posteriori (MAP). Given a problem with K classes {C1... CK} with so-called prior probabilities P(C1)...P(CK), we can assign the class label c to an unknown example with features x = (x1 , . . . , xN) such that c = arg max cP (C = ck x1, . . . ,xN), that is choose the class with the maximum aposterior probability given the observed data. This aposterior probability can be formulated, using Bayes theorem, as follows: P (C = ckx1, . . ,xN) = P (C = c) P (x1, . . . , xNkC = c) P (x1, . . . , xN) As the denominator is the same for all clases, it can be dropped from the comparison. Now, we should compute the so-called class conditional probabilities of the features given the available classes. This can be quite difficult taking into account the dependencies between features. Then aive bayes approach is to assume class conditional independence i.e. x1, . . . , x Nare independent given the class. This simplifies the numerator to be P (C = c) P (x1kC = c) . . . P (xNkCs = c), and then choosing the class c that maximizes this value over all the classes c = 1, . . . ,K.

D) Support Vector Machine

Support Vector Machines are among the most robust and successful classification Algorithms. It is a new classification method for both linear and nonlinear Data. It uses an online ar mapping to transform the original training data into a higher dimension. With the new dimension, it searches for the linear optimal separating hyper plane (i.e., “decision boundary”). With an appropriate non linear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyper plane. SVM finds this hyper plane using support vectors (“essential” training tuples) and margins (defined by the support vectors)[15].

V. WEB PAGE CLASSIFICATION TECHNIQUES

Web pages can be classified into the following categories:

a. Manual classification
b. Clustering approaches
c. META tags based categorization

d. Text content based categorization
e. Link and content analysis

In addition to these, document structure based approach has also been used for classifying web pages [8].

VI. WEB DOCUMENT CLASSIFICATION METHODS

Web Page Classification contains three phases. They are

(a) Feature Extraction Phase : In this phase features are extracted from the training dataset
(b) Feature Selection Phase: In this phase features are selected from the web pages for each and every category.
(c) Classification Phase: In this phase different classification techniques applied like naïve bayes, SVM, lazy learners, k-nearest neighbor (KNN) etc.

VII. CONCLUSION

This Paper represent various web mining consept,In that it specially focus on the Web Content Mining. It also shows the different Web Page Classification Categories and they works. It also gives the idea how information retried for the web. For the future work, improve the result of web page classification techniques.

REFERENCES

[1] J.Krutil,M.Kudelka and V.Snasel,"Web Page Classification based on Schema.org Collection", In: Fourth International Conference on Computational Aspects of Social Networks(CASoN),2012

[2] Sarac,E.;Ozel,S.A.,"Web Page Classification Using Firefly Optimization",In:Innovations in Intelligent Systems and Applications(INISTA),2013 IEEE International Symposium, pages 1-5

[3] Wang Zhixing and Chen Shaohong,"Web Page Classification based on Semi-supervised Naive Bayesian EM Algorithm",In:IEEE International Conference on Communication Software and Networks(ICCSN),2011 pages,242-245

[4] Yusuf,L.M.and Othman,M.S.,Salim,J.,"Web Classification using Extraction and Machine Learning Techniques", In: Information Technology(ITSim)2010 International Symposium in (volume:2),

[5] Tian Xia,Yanmei Chai,Tong Wang,"Improving SVM on Web Content Classification by Document Formulation", In: 7th International Conference on Computer Science & Education(ICCSE 2012)

[6] D.Navadiay,M.Parikh,R.Patel,"Constructure Based Web Page Classification", International Journal of Computer Science and Management Research,Vol 2,Issue 6, June 2013

[7] M.IndraDevi,Dr.R.Rajaram,K.Selvakuberan,"Automatic Web Page Classification by Combining Feature Selection Techniques and Lazy Learners",In: International Conference on Computational and Multimedia Applications(ICCIMA),2007

[8] A.Asirvatham,K.Ravi,"Web Page Categorization based on Document Structure",www.iiit.ac.in/~arul/paper.pdf

[9] ZHI-MING XU,XIN-BO GAO,MENG LEI,"WEB SITE CLASSIFICATION BASED ON KEY RESOURCES",In: Proceedings of the 8th International Conference on Machine Learning and Cybernetics,2009

[10] Selma Ayse Ozel,"A Web page Classification System Based on a genetic algorithm using tagged-terms as feature",In: Journal On Expert System Applications 38(2011)3407-3415

[11] V.Fernandez,R.Unanue,S.Herranz,A.Rubio,"NaıveBayesWebPageClassificationwithHTMLMark-UpEnrichment",In Proceedings of the International Multi-Conference on computing in the Global Information Technology,2006.

[12] S.Meher,S.Pal,S.dutta,"Granular Computing Models in the Classification of Web Content Data",In International Conferences on Web Intelligence & Intelligent Agent Technology,2012.

[13] S.Balan, " A study of various Techniques of Web Content Mining Research Issues and Tools",In International Journal of Innovative Research and Studies,Vol-2,Issue-5,May-2013.pp508-517

[14] Cooley R., Mobasher B, Srivastava J,"Web mining: information and pattern discovery on the World Wide Web",.In Proceedings of Ninth IEEE International Conference.Nov-1997.Pp558-567.

[15] D.Navadiya,R.Patel,"Web Content Mining Techniques-A Comprehensive Survey", International Journal of Engineering Research & Technology(IJERT),Vol.1 Issue 10,December-2012,Pp1-6.