IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE)
e-ISSN: 2278-1684,p-ISSN: 2320-334X, Volume 13, Issue 3 Ver. VI (May- Jun. 2016), PP 01-09
www.iosrjournals.org

Web Pages Classification Using Rule-Based System

Zinah A. Abbas[1], Prof .Dr.Shawkat K. Guirgui[2], Dr .LaithR. Fleih[3], Magda M. Madbouly[4]

[1]*Department of computer Science Ministry of Labor & Social Affairs Iraq-Baghdad*
[2]*Department of Information Technology Institute of Graduate Studies and Research Alexandria University*
[3]*Dept. Computer science College ofScienceCihan University/Iraq/Kurdistan*
[4]*Department of Information Technology Institute of Graduate Studies and Research Alexandria University*

Abstract: *The Web is the largest collection of electronically accessible documents which make the richest source of information in the world. The problem with the Web is that this information is not well structured and organized so that it would be easily retrieved. Web page classification used for managing and extract relevant information from Web content and in order to effectively use the knowledge available on the Web. In this dissertation a rule base system used to contract the system classifier for solve Web online classification by assigning each scanned HTML document to their class. The aim of this thesis is to design and implement HTML document classification system that is able to classify the HTML documents according to their class (category). The proposed system is designed to solve and improve web page classification problem using Rule-Based Classifier that check the HTML content of each entered URL address for system rule occurrences. The Proposed system enhances other web page classification system by making the system work online.*

Keywords: *Web page classification, Rule- based system, Text classification, and HTML classification.*

I. Introduction

The world is moving across the internet. 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. 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 [1].

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. Web page classification is a web mining area, using this method we can identify web pages, web page classification retrieves web pages based on content and structure of web page.

Web page Classification is a process where one page is appended to one or more directories which is predefined in advance [2]. 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 [3].

The categorization techniques can be classified into the following broad categories [4]:-

a- Categorization by domain experts
b- Clustering approaches
c- Meta tags based approach
d- Text content based categorization
e- Link and Content Analysis

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 [2].

There are several applications for the classification of the web page and they are [2]: -

1-Constructing, maintaining or expanding web directories (web hierarchies)
2-Improving quality of search results
3-Helping question answering systems
4-Building efficient focused crawlers or vertical search engines
5-Web content filtering
6-Assisted web browsing
7-Knowledge base construction

In addition to these, document structure based approach has also been used for classifying web pages. The two common tasks through which useful information can be mined from Web are Clustering and Classification [2].

II. Classification System

Considered World Wide Web is the largest database in the Universe which is mostly understandable by human users and not by machines. It lacks the existence of a semantic structure which maintains interdependency of its components. Presently, search on web is keyword based i.e., information is retrieved on the basis of text search of all available matching URL's / hyperlinks. This may result in the presentation of irrelevant information to the user. In the current web, resources are accessible through hyperlinks to web content spread throughout the world [5].

The general problem of web page classification can be further divided into multiple sub-problems such as subject classification, functional classification, sentiment classification, and other types of classification. Subject classification is concerned about the subject or topic of a web page [6]. There are many machine learning algorithms used for web page classification are [7]:-

2.1 Association rule mining: The formal statement of Association rule mining problem was initially specified by Agrawal. Association rule mining discovers the frequent patterns among the item sets. It aims to extract interesting associations, frequent patterns, and correlations among sets of items in the data repositories. And the main sub problem can be two folded into candidate large item sets generation process and frequent item sets generation process. Those item sets whose support exceeds the support threshold called as large or frequent item sets, those item sets that are expected to be large or frequent are known candidate item sets. An efficient model has classification rules with high confidence and large support [8].

2.2 Naive Bayes: Bayesian Network consists of two components. First component is mainly a directed acyclic graph (DAG) in which the nodes in the graph are called the random variables and the edges between the nodes or random variables represents the probabilistic dependencies among the corresponding random variables. Second component is a set of parameters that describe the conditional probability of each variable given its parents. The conditional dependencies in the graph are estimated by statistical and computational methods. Thus the Bayesian Network combines the properties of computer science and statistics [9].

2.3 Support Vector Machine (SVM): SVMs are widely regarded as state-of-the-art models for binary classification of high dimensional data. SVMs are trained to maximize the margin of correct classification, and the resulting decision boundaries are robust to slight perturbations of the feature vectors, thereby providing a hedge against over fitting. The superior generalization abilities of SVMs have been borne out by both theoretical studies and experimental successes. The required optimization can be formulated as an instance of quadratic programming, a problem for which many efficient solvers have been developed. Have been experimented with both linear and radial basis function (RBF) kernels [10].

2.4 Logistic Regression: This is a simple parametric model for binary classification where examples are classified based on their distance from a hyper plane decision boundary. We included $\ell 1$-regularized logistic regression for its potential advantages over Naive Bayes and SVMs in particular domain. Unlike Naive Bayes classification, the parameters in logistic regression are estimated by optimizing an objective function that closely tracks the error rate. Unlike SVMs, $\ell 1$-regularized logistic regression is especially well suited to domains with large numbers of irrelevant features. (In large numbers, such features can drown out the similarities between related examples that SVMs expect to be measured by the kernel function.) Finally, because $\ell 1$-regularized logistic regression encourages sparse solutions, the resulting models often have decision rules that are easier to interpret in terms of relevant and irrelevant features [10].

2.5 Decision trees: Was used J48 to build standalone decision trees and to construct bagging ensembles. Although ensemble classifiers are more complex and thus demand more time, they are often beneficial in terms of tradeoff between additional time and improved performance [11].

2.6 The Flexi Rank algorithm operates on a set of web pages returned by a web crawler and gives a ranking of the pages as output. A web page ranking algorithm it works based on syntactic classification of web pages. The mainly consists of three steps: select some properties of web pages based on user's demand, measure them, and give different weightage to each property during ranking for different types of pages. The existence of syntactic classification is supported by running fuzzy c-means algorithm and neural network

classifier on a set of web pages. It has been demonstrated that, for different types of pages, the same query string has produced different page ranking [12].

III. Web Page Classification

Web mining is the data mining technique that automatically discovers/extracts the information from web documents. It is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. The term Web Data Mining is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest. Mining techniques are in detail, results and comparison to extract necessary information effectively and efficiently [13].

Well as web mining technique is used to fetch knowledge form Web data. Web mining can be broadly defined as the search and measure of useful information from the World Wide Web. Web Mining is the extraction of useful patterns and implicit information related to the World Wide Web [14].
Web mining it could be done in three different ways as shown in Figure 1 [15]:-

1. Web Content Mining: It focuses on extracting knowledge from the contents or their descriptions. It involves techniques for summarizing, classification and clustering of the web contents. It can provide useful and interesting patterns about user needs and contribution behavior [16]. It is related to text mining because much of the web contents are text based. Text mining focuses on unstructured texts. Web content mining is semi-structured nature of the web. Technologies used in web content mining are NLP, IR [17].

2. Web structure mining: - The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting two related pages. It is the process of discovering Information from the Web. It is used for finding information about the web pages and inference on Hyperlink. The Web consists not only of pages, but also of hyperlinks pointing from one page to another. It discovers the link structure of the hyperlinks at the inter-document level and to generate structural summary about the Website and Web page. It is used for retrieving pages that are not only relevant but are also of high quality, or authoritative on the topic [14].

3. Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a web site [5].
   Adaptive arithmetic coder before being written to the compressed file [2].

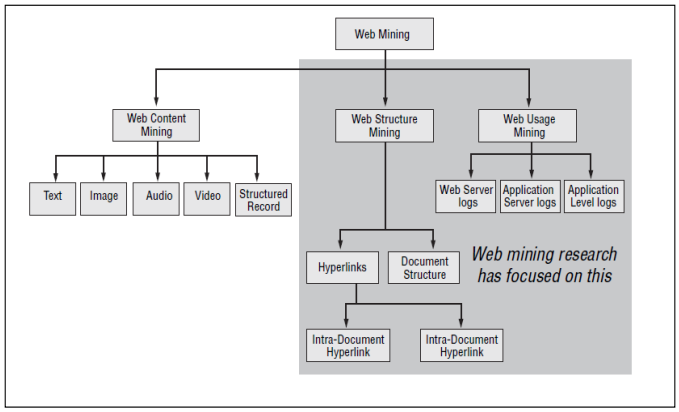

Fig.1: Structure of Web mining [15]

IV. Rule-Based Classification

Association rule mining is a technique in data mining which is used to mine different association rules from the given database [18], and considered one of the most important tasks in data mining [19].

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. [18]. Association rule mining has gained a lot of attention in the field of research. This technique aims at finding dependences among the attributes in a database. Therefore, it has found applicability in a lot of fields [20], such as telecommunication networks, risk and market management, inventory control, medical diagnosis/drug testing etc. [21]. Many algorithms have been proposed

that mine association rules from a database based on the minsup and minconf specified by the user [20]. The different areas rules play an important role in, it is coarsely distinguished between application areas of agents with rules and multi-agent system construction aspects. Interesting application areas of rules and agents that include rather traditional fields like parallel and distributed rule-based systems, service oriented architecture, grid and peer-to-peer computing as well as upcoming new trends such as cloud computing, rule based wireless networks and complex event processing scenarios [22], and cloud computing and data mining have become famous phenomena in the current application of information technology [23].

And there are two important basic measures for association rules, support(s) and confidence(c). The two thresholds are called minimal support and minimal confidence respectively [24].

A. Support(s) is defined as the proportion of records that contain X □ Y to the overall records in the database. The amount for each item is augmented by one, whenever the item is crossed over in different transaction in database during the course of the scanning [8]. It is inefficient in case of huge mining problems. In the classical association rule mining algorithms users have to specify the minimum support for the given dataset upon which the association rule mining algorithm will work. But it is very much possible that the user sets a wrong minimum support value which can hamper the generation of association rules. And also there is the fact that the setting of this minimum support is also not an easy task. If minimum support is set to a wrong value then there is a big possibility of combinatorial blow up of huge number of association rule within which many association rules will not be interesting, well as considered very a difficult task of setting an accurate minimum support value manually [25].

B. Confidence(c) is defined as the proportion of the number of transactions that contain X □ Y to the overall records that contain X, where, if the ratio outperforms the threshold of confidence, an association rule X □ Y can be generated [8]. Considered confidence level are key issues for particular data mining tasks[24], And that the confidence value of the rule should increase by increasing the number of antecedent items. As well the confidence values of association rules are calculated and the ones which have confidence value equal to or greater than the predefined confidence value are the final output of the algorithm.

V. Text Classification

Automatic classification of Web documents has become important for effective retrieval. As one of the essential techniques for Web mining, Web document classification has been studied extensively [26]. Within the field of machine learning, classification is the problem of training a learning algorithm on a set of labeled examples belonging to two or more classes so that, when given a new unlabeled instance, the algorithm may assign the correct label. This is referred to as a supervised learning problem because the training examples must be labeled—usually by a human. An unsupervised learning problem, by contrast, takes unlabeled instances and derives both a set of labels and a procedure for assigning one or more of those labels to each instance. The unsupervised learning equivalent of classification is called categorization [27]. Text taxonomy is a kind of "supervised" learning where the categories are known beforehand and determined in advance for each training document. Documents pre-processing allows an efficient data manipulation and representation [28].The text taxonomy process is illustrated in the figure-2 [29].

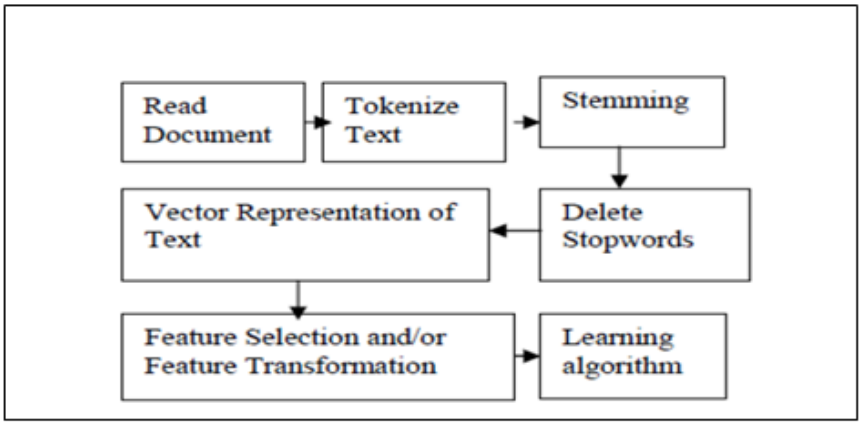

Fig.2: Taxonomy of the Text Classification Process [29].

We briefly describe as following [30]:-

1. Read Document step: at first all of documents are read.
2. Tokenize text step: in this step the text is broken into tokens, meaningful words, terms, phrases, symbols or elements which is called Tokenization.
3. Stemming: step: the root of words is transformed into an original form.
4. Stop words step: words such as in, this, a, an, the, with and etc. are removed.
5. Vector Representation of Text: In this step, an algebraic model is defined to represent text documents as a vector. Because the main goal of feature selection methods is to reduce dimensionality.
6. Feature Selection and/or Feature Transformation: In this step we reduce the dimensions of datasets using feature selection methods by removing the features not related to classification. After documents feature selection, according to the flexibility, we can use machine learning algorithms such as Genetic algorithm, Neural Network, Rule Induction, Fuzzy Decision Tree, SVM, K-NN (Nearest Neighbor) algorithm, Lsa, Rocchio algorithm and Naïve Bayesian.

VI. Propose System

In this section we will explain the process for each interface of the system interfaces we have done. Each interface contains several key parts, with a specific job of each part to do. There are some key parts containing certain branches where each branch does its own job. The entire main and branch parts help the user to know the classification of pages on the Internet. The classification in our system depends on the presence of the URL which is the most significant part of the system, when introduced in our system we'll get the rating of the page, either in the case it is absent, the system is no more working.

The Main interface: This interface is considered the first of our system, it contains three main parts; categories manager, classification and URL manager.

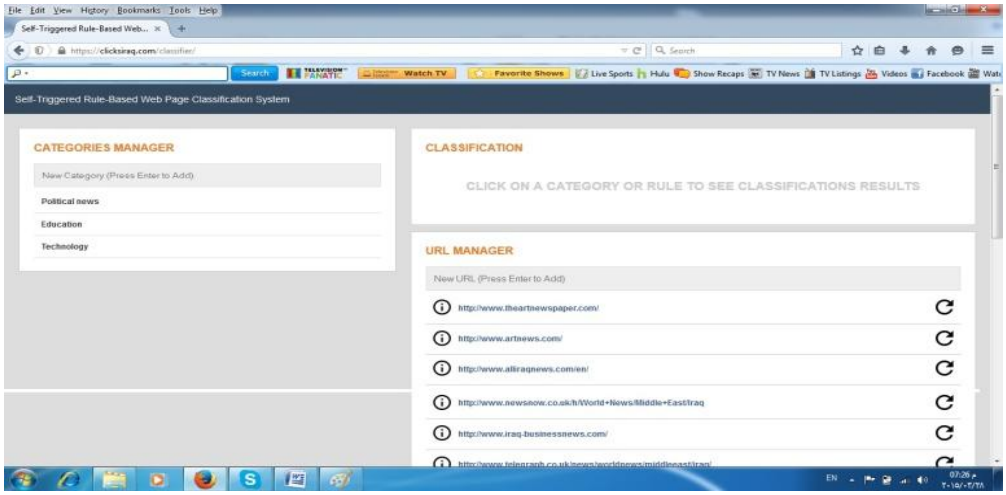

Fig.3: The Main interface

Categories manager; in this section we introduce the main class, which contains one rule or a number of them. Based on the category and its rules the internet pages are classified, the category should be correct containing the correct rules. If there was a mistake in the class or the rule, a system would make not any practical classification of Internet pages. There is a branching-part of the manager categories; a New Category (Press Enter to Add). In this section we would set a master class or more, the more the number of categories or rules, the more accurate classification we get. There are three main categories in this interface, Political news, Education and Technology, each one of which contains a number of rules we will describe later.

Classification is a part of the main parts of the main interface, where we would find a sentence asking the user to click on a category or rule to see the results of the classification, there is a difference between clicking on the rule or category. The difference is when the user clicks on a specific category system which brings the entire URL containing the category and its rules, either when you click on one of a certain class rules, the system will bring the entire URL that contain this rule only.

URL manager: is an important part of the main interface, it is the part of the process that manages the URL of when the introduction of a new URL, as well as the introduction of each new URL in the library manager, those will be subject to a classification process. When entering a correct URL, the system will conduct the classification process and show its own results, as in the case of the new fact that the URL is incorrect; the

system displays a message saying that this URL is not show able. Two signs will show when entering a new URL, one of them will appear to the right of the URL and the other on its left. When clicking on the tag located on the right, the system will start a new page to scan the Internet, whereas the mark to the left of the URL will show the results of the classification process. Clicking results on these signs are changeable due to the change of the Internet page, but if there is no change in the internet page, the results will remain the same as they are.

Add a new URL for the system: There are three main parts in the main interface of the system; one of these parts is the URL manager which works on the introduction and conduct of each new URL into the system, so the results of the classification of the Internet pages can be reached through this process.

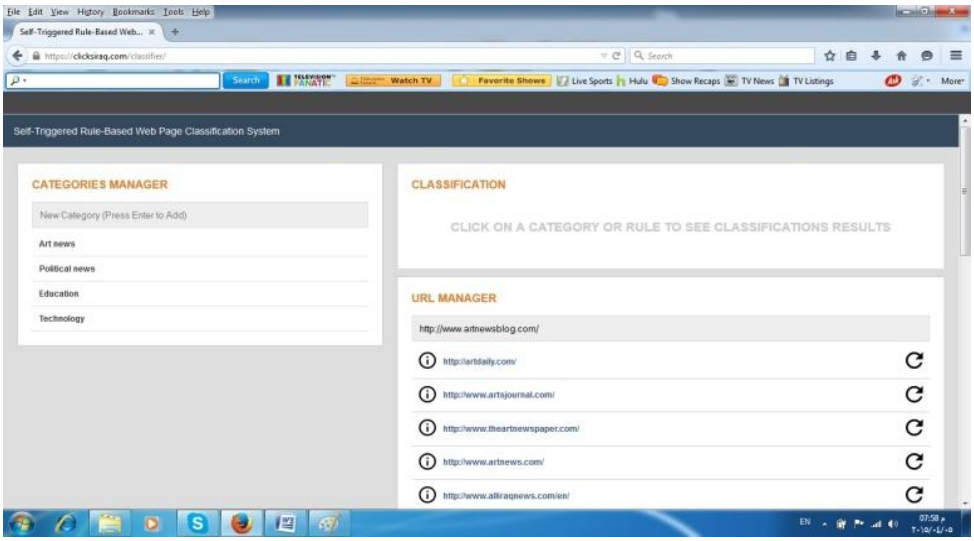

Fig.4: Add a new URL for the system

In this interface, we have added a new URL in an allocated place under the phrase "URL MANAGER", then by clicking on enter it'll take its sequence in the library manager.

Enter the URL of the new system: The URL is entered after entering the new URL and clicking on enter in the previous interface, a sign on the right and left sides of the URL appeared, each of the two signs would have a particular job, we'll describe later.

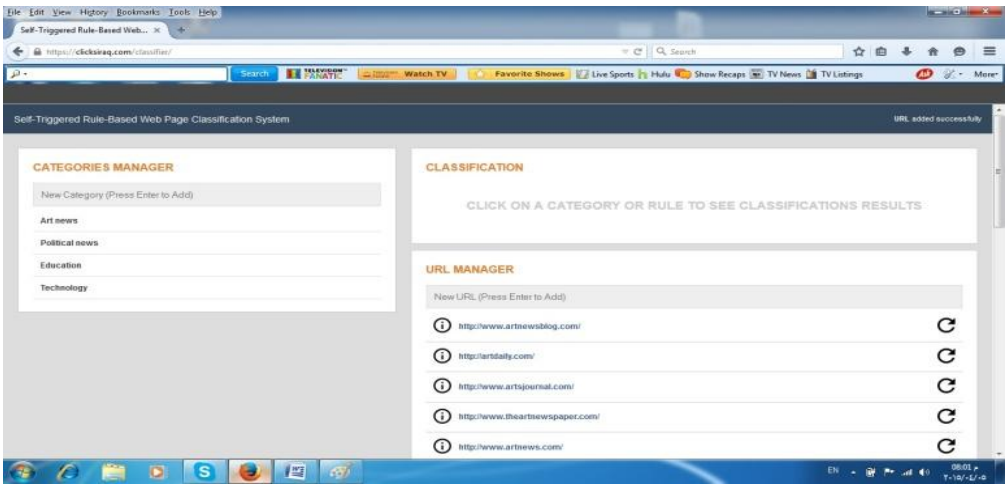

Fig: 5 enter the URL of the new system

A message notifying that the URL added successfully had shown, indicating that the new URL which has been entered for the system is also true.

URL SCANNER: In the previous interface we've mentioned that there are two signs would appear when entering a new URL into the system. In this interface we will explain the sign to the right of the URL that was given the symbol (↻). When clicking this sign a URL SCANNER interface with two words; Cancel and Start

Scan will appear. So, if we wanted to conduct a scan we'll need to click on Start Scan, otherwise, in the case we're willing to cancel the order Scan we'll need to click on Cancel. In the case of clicking on the Start Scan, the system will show the original text of the Internet page to the left, then start making two processes; the first is Stop Word which works on removing the entire non-useful words in the classification process, the second process is the one called stemming, which works on restoring each word to its original root. To reduce the size of the text we've conducted these operations so as to rapidly conduct the classification process to get the best results. After these two processes are done, the text will appear in a smaller size beside the original text, the results will also appear to the right of the interface, and underneath the number of words repetition we've already entered being rules of main categories. In this interface, the word Music appeared twice, considered one of the rules of Category Art News, The other rules shown no signs of repetition because they were not existed within the original text of the new URL we've entered.

Fig: 6 URL SCANNER

URL Classification: After all the previous steps are done, the final results of the process of the new URL classification would appear in this interface. In the previous interface, we've explained the processes taking place when clicking on the sign to the right of the URL, but in present interface we will click on the sign to the left of the URL, which had been given the symbol (ⓘ).

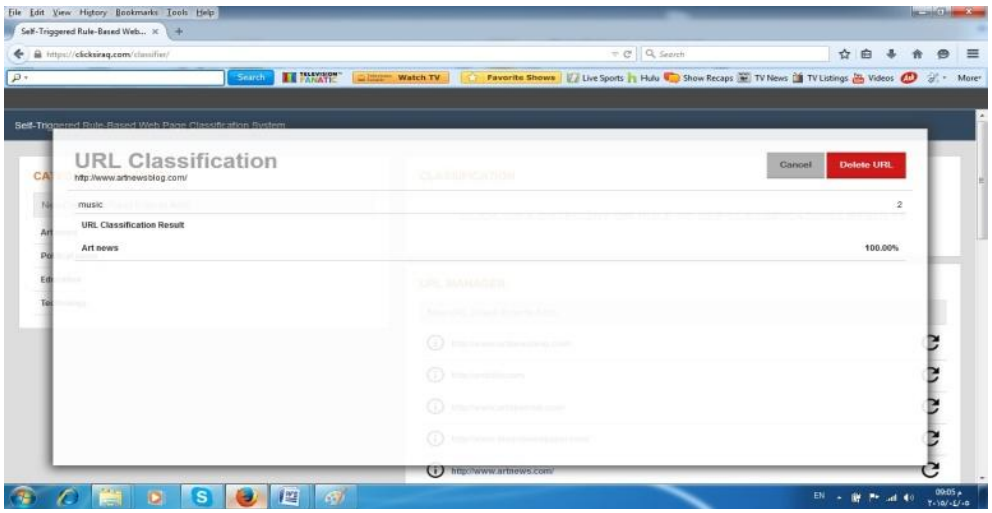

Fig: 7 URL Classification

After clicking, two words on the right side will appear; Cancel and Delete URL, so if we wanted to go back to the main interface, we need to click on Cancel, but, if the user wishes to permanently remove the URL from the system, he'll need to click on Delete URL.

The URL, That a classification process was acquired with a rule underneath named , Music, would appear, the system would also notify us that this rule is repeated twice and the results of the new URL

classification is 100%the category Art news, this means that the URL lacks rating of the other categories existing in the system.

VII. Conclusion

We may consider it efficiently performing the processes of classifying the data. Therefore, we may conclude that the results of the classification are proper and correct. We can also consider the process of using the system is a very easy one because our system is UN complicated and easy to use. The system restores all the data on the Internet page to get to the results of rating this page when an infinite number of the URL can be entered into the system where each URL has its own rating results.

There are several characteristics in our system, the first of which is the rating of the data available on the net pages in the form of Online, In other words, the system is sensing the data on the net pages.

The second characteristic for which the system was built; is the classification of the results. We mean that the system displays the classification results of any page of the net in an easy, clear and unambiguous way for the user. In addition, through these results, we get a private or important data about the classification of this page.

The third characteristic is the speed of the system, by which we mean that this system is fetching the URL from a very large database available on the Internet, It also processes a number of operations, including scanning, stop word, stemming and others. Eventually it gives the ranking results of the page, all these operations carried out by the system very quickly not to exceed one minute.

At the end of this thesis we conclude that this system has many significant characteristics that help the system user to take advantage of it in many areas, and the view of the above we can consider that more than 98% accuracy of this system is where the classification of pages on the Internet are true and very accurate.

We also have three proposals to develop our system; the first is that the system do classify the HTML documents only and can in the future be used for the classification of new change able types such as the images, audio and multimedia.The second proposal; is that we have done the classification process using the Rule-Based system, so we can develop our system using new methods such as artificial neural network.The third and last one; is that the system can be developed in order to be able of providing suggestions on how smart search the internet.

References

[1]. T. Xia, Y. Chai, "Improving SVM on Web Content Classification by Document Formulation", In Proceedings of the 7th International Conference on Computer Science & Education (ICCSE '12), Australia, 14-17 July 2012, pp. 110-113.

[2]. S. D. Vaghela, P. Patel, "Web page Classification Techniques-A Comprehensive Survey", International Journal of Engineering Development and Research (IJEDR), Dec. 2014, Vol.1, Issue 3, pp.229 – 232.

[3]. T. Slimani, A. Lazzez, "Efficient Analysis of Pattern and Association Rule Mining Approaches", International Journal of Information Technology and Computer Science (IJITCS), Feb. 2014, vol. 6, No. 3, pp.70-81.

[4]. A. P. Asirvatham, K. K. Ravi, "Web Page Categorization based on Document Structure", International Institute of Information Technology, Hyderabad, India 500019, 2001.

[5]. V. Jain and Dr. S. V. A. V. Prasad, "Ontology Based Information Retrieval Model in Semantic Web: A Review", Ontology Based Information Retrieval Model in Semantic Web: A Review, Vol. 4, Issue 8, August 2014, pp. 837-842.

[6]. P. Manchanda, S. Gupta and K. K. Bhatia, "On The Automated Classification of Web Pages Using Artificial Neural Network", IOSR Journal of Computer Engineering IOSRJCE, Vol. 4, Issue 1, Sep-Oct. 2012, PP. 20-25.

[7]. M. I. Devi, Dr. R. Rajaram and K. Selvakuberan, "Automatic Web Page Classification by Combining Feature Selection Techniques and Lazy Learners", International Conference on Computational Intelligence and Multimedia Applications (ICCIMA -07), pp. 33-37, Sivakasi, Tamil Nadu, 13-15 Dec. 2007, Vol. 2.

[8]. T. Karthikeyanand N. Ravikumar, "A Survey on Association Rule Mining", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2014, pp. 5223-5227.

[9]. Ms S. Vijayarani and Ms M. Muthulakshmi, "Comparative Analysis of Bayes and Lazy Classification Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, Aug. 2013, pp. 3118-3124.

[10]. Justin Ma, L. K. Saul, S. Savage and Geoffrey M. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs", Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 1245-1254, Paris, France, June 28–July 1, 2009.

[11]. VedranaVidulin, MitjaLu˙strek and Matja˘z Gams, "Training a Genre Classifier for Automatic Classification of Web Pages", Journal of Computing and Information Technology - CIT 15, Vol. 15, Issue 4, Dec. 2007, pp. 305–311.

[12]. D. Mukhopadhyay and P. Biswas, "FlexiRank: An Algorithm Offering Flexibility and Accuracy for Ranking the Web Pages", Proceedings of the ICDCIT 2005, pp. 308-313, Bhubaneswar, India, 22-25 Dec. 2005; LNCS 3816, Springer-Verlag, Berlin, Germany 2005.

[13]. N. R. Kapadia and K. Patel, "Web content mining techniques – a comprehensive survey", International Journal of Research in Engineering & Applied Sciences, Vol. 2, Issue 2, Feb. 2012, pp. 1869-1877.

[14]. S. Sharama, "Web Mining", International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459), Vol. 2, Issue 4, April 2012, pp. 269-271.

[15]. T. S. srivasta, P. D. desikan and V. K. kumar, "Foundations and Advances in Data Mining", Springer Berlin Heidelberg, Springer-Verlag Berlin/Heidelberg, 2005.

[16]. A. Singh, "Agent Based Framework for Semantic Web Content Mining", International Journal of Advancements in Technology, Vol. 3, No. 2, April 2012, pp. 108-113.
[17]. S. Lingwal and B. Gupta, "A Comparative Study Of Different Approaches For Improving Search Engine Performance", International Journal of Emerging Trends & Technology in Computer Sciench (IJETTCS), Vol. 1, Issue 3, Sep. – Oct. 2012, pp. 123-132.
[18]. I. Chandrakar and A. M. Kirthima, "A Survey on Association Rule Mining Algorithms", International Journal of Mathematics and Computer Research, Vol. 1, Issue 10, Nov. 2013, pp. 270-272.
[19]. J. K. Jain, N. Tiwari and M. Ramaiya, "A Survey: On Association Rule Mining", International Journal of Engineering Research and Applications (IJERA), Vol. 3, Issue 1, Jan.–Feb. 2013, pp. 2065-2069.
[20]. N. Sharma1and Dr. C. K. Verma, "Association Rule Mining: An Overview", IJCSC, Vol. 5, No. 1, Mar. 2014, pp. 10-15.
[21]. G. Kaur, "Association Rule Mining: A Survey", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5, No. 2, 2014, pp. 2320-2324.
[22]. N. Bassiliades, G. GovernatoriandA. Paschke, "Rule-Based Reasoning, Programming, and Applications", Springer Berlin, 5th International Symposium, Rule ML 2011.
[23]. V. Jain and Dr. S. V. A. V. Prasad, "MINING IN ONTOLOGY WITH MULTI AGENT
[24]. SYSTEM IN SEMANTIC WEB: A NOVEL APPROACH", the International Journal of Multimedia & Its Applications (IJMA), Vol. 6, No. 5, Oct. 2014,pp. 45-54.
[25]. S. Kotsiantis and D. Kanellopoulos, "Association Rules Mining: A Recent Overview", GESTS International Transactions on Computer Science and Engineering, Vol. 32, No. 1, 2006, pp. 71-82.
[26]. A. Roy and R. Chatterjee, "A Survey on Fuzzy Association Rule Mining Methodologies", IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 15, Issue 6, Nov. - Dec. 2013, PP. 1-8.
[27]. Q. Duan, D. Miao and M. Chen, "Rough Sets, Fuzzy Sets, Data Mining and Granular Computing", Springer Berlin Heidelberg, Springer Berlin Heidelberg, 2007.
[28]. Shawn C. Tice, "Classification of Web Pages in Yioop with Active Learning", May 2013, Master's Projects.Paper 297, San Jose State University.
[29]. Mr. Suresh G S and Mrs. SharayuPradeep, "Realization of computerized text taxonomy through a supervised learning system", International Journal of Engineering and Computer Science ISSN: 2319-7242, Vol. 3, Issue 5, May 2014, pp. 6022-6026.
[30]. D. B. Dasari and Dr. VenuGopalaRao. K, "Text Categorization and Machine Learning Methods: Current State of the Art", Global Journal of Computer Science and Technology, Vol. 12, No. 11, 2012, pp. 37-46.
[31]. N. VasfiSisi1 and M. R. F. Derakhshi, "Text Classification with Machine Learning Algorithms", Journal of Basic and Applied Scientific Research, 3(1s), July 7-2013, pp. 31-35.