© 2017 IJSRSET | Volume 3 | Issue 8 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Web Page Categorization through Data Mining Classification Techniques on URL Information

R. GeethaRamani[1], P. Revathy[*2]
[1]Department of Information Science and Technology, College of Engineering, Anna University, Chennai, India
[*2]Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, India

ABSTRACT

The usage of web is increasing exponentially every day. The huge deluge of information owing to the web access can be mined to reveal interesting patterns. Web mining is gaining popularity in the recent times. In this paper web page categorization is attempted. Classification of web pages can provide useful information with regard to advertising and recommendations. Earlier techniques have utilised context, source and URL based information for classification of web pages. In this work, web page classification is performed through URL information. The proposed methodology involves data pre-processing, feature vector formulation, C4.5 classification, performance evaluation and prediction of class of web page. The experimentation has been carried out with NASA log dataset. Various classifiers have been utilised and C4.5 proves to yield the best possible results achieving an accuracy of 99.80% using 3 - fold cross validation. The results obtained justify the performance of the proposed methodology.

Keywords: Web Page Categorization, Web Usage Mining, C4.5, NASA Dataset, URL Information

I. INTRODUCTION

Data mining [1] is the science of discovering interesting patterns from the huge deluge of data. It finds its application in many fields such as web [2], medicine [3], entertainment [4] etc. The emerging of web mining [5] emphasizes the extent to which data mining is fruitful to the web domain. Web mining refers to application of data mining methods to unravel interesting patterns from the World Wide Web (WWW). Web can be viewed as an ocean of different kinds of data from various sources. Hence, various types of mining through a variety of techniques can be performed to discover useful information. In this context, web mining can be broadly categorised into three types namely web usage mining, web content mining and web structure mining [6]. Web usage mining [7] deals with mining the web usage patterns of users based on their navigation behaviour. Web content mining [8] refers to mining the content of the web pages in the view of bringing to light the non-trivial information. Web structure mining [9] is related to identifying patterns from structure of the websites.

In this work, web usage mining is placed the main focus. Web usage mining primarily concentrates on analysing the usage patterns of web pages by the users. For this purpose, various mining techniques namely classification [10], clustering [11] and association rule mining [12] etc., can be adopted. Classification requires labelled training data for its operation while clustering do not require labelled training data. Classification has proved to be more effective than clustering in many applications. In this paper, categorisation of web pages through classification is attempted. This can be utilised to provide apt recommendations to the users when they view a particular category of webpage.

The remaining of the paper is organized as follows: Section 2 presents the literature survey related to web page categorization; Section 3 describes the proposed methodology; Section 4 discusses the associated results and Section 5 concludes the paper.

II. LITERATURE SURVEY

The existing works related to web page categorization is concisely presented here.

In 2010, data mining techniques have been operated on Chinese web pages in the view of classifying them [13]. K-Nearest Neighbour, Support Vector Machine and Adaptive Resonance Association Map have been evaluated in the context of macro and micro-level classification of web pages. It has been observed that Adaptive Resonance Association Map and K-Nearest Neighbour performed superior to Support Vector Machines.

In 2003, web page classification has been performed through adaptive ontology [14]. Initially representative terms are recognized through the term frequency and product frequency. Then, information gain is computed to prioritise it for classification. Classification rules are then formulated through these selected terms. Then, the rules will be able to classify the web pages into any one of the category as in the domain ontology. The system achieves an accuracy of 95.29%.

In 2006, classification of web pages has been attempted through Artificial Neural Networks [15]. The methodology comprises of three steps involving extraction of features from the source of the web page, assigning appropriate input values to the network and prediction of class through back propagation neural network. The method yields an accuracy of 70.67% on classifying randomly selected 300 web pages into eight categories.

In 2011, web page classification is tried through optimum number of features [16]. The methodology involves pre-processing, feature set extraction, Correlation based feature selection, building of learning model using C4.5, extraction of final set of features from the pruned decision tree, building of learning models using the final set of features. Decision Tree, K-Nearest Neighbour, OneR, Multi Layer Perceptron and Radial Basis Function have been evaluated for building of learning model out of which Radial Basis Function, K Nearest Neighbour and Multi-Layer Perceptron achieved the best performance yielding an accuracy of 99%. The feature extraction methodology proves to improve the accuracy of all the considered classifiers.

In 2013, web page categorization through Multi-Layer Perceptron has been put forth [17]. Initially, the HTML and URL features of the web pages are extracted followed by selection of optimal features through feature reduction procedure. Then, Multi-Layer Perceptron techniques is utilised to build the learning model on the reduced features. The model achieves an accuracy of 96.60% on classifying 210 randomly selected web pages into six categories.

Having provided a glimpse on the earlier works towards web page classification, the proposed approach attempts to classify the web pages only based on the URL information. The proposed methodology is described in the following section.

III. PROPOSED METHODOLOGY

Web page classification will be very useful in the context of advertising and recommendations. The proposed approach attempts to categorize the web pages based on the information available in their URL. The proposed methodology is depicted in Figure1.

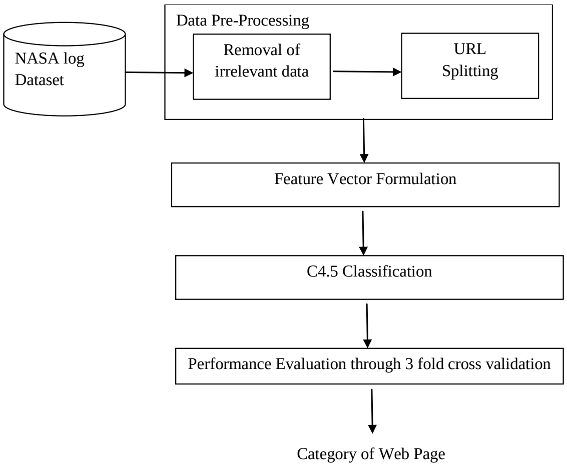
Data Pre-Processing
NASA log Dataset
Removal of irrelevant data
URL Splitting
Feature Vector Formulation
C4.5 Classification
Performance Evaluation through 3 fold cross validation
Category of Web Page

Figure 1. Proposed Framework

The proposed framework incorporates data pre-processing, feature vector formulation, application of classification algorithms, performance evaluation to evolve the efficient rules for classification of web pages. Every step-in the proposed framework along with the data used for its evaluation is presented in detail subsequently.

A. Dataset Description

The dataset used for training and testing the learning model is described here. The dataset used for experimentation is the NASA log data [18] [19]. It has been downloaded from the NASA Kennedy space center server in Florida. The data log contains the entries for 31 days from 1st July, 1995 to 31st July, 1995.

It is composed of more than 1000000 entries. The entries constitute (i) IP address of the user (ii) Username and Password (iii) Date and time of access (iv) Request method which contains value (GET, POST, and HEAD), (v) URL of the requested web page (vi) Status code and (vii) Bytes of data transferred. A sample log entry from NASA dataset is shown in Figure2.

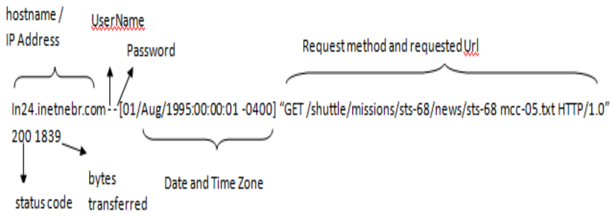
hostname /
IP Address
UserName
Password
Request method and requested Url
In24.inetnebr.com - - [01/Aug/1995:00:00:01 -0400] "GET /shuttle/missions/sts-68/news/sts-68 mcc-05.txt HTTP/1.0"
200 1839
bytes
transferred
Date and Time Zone
status code

Figure 2. Sample Log Format of NASA Dataset

For the current work, only entries of the first day are considered. This comprises of 64715 entries. The further sub-sections present the processing done on the data to arrive at web page classification.

B. Data Pre-Processing

The raw data obtained from the NASA log is pre-processed to make it suitable for subsequent processing. This stage involves irrelevant data removal and URL splitting. Initially, the irrelevant data is removed. The process removes the instances with error codes and hence retains only those instances which have a status code of 200. Then all the multimedia content is eliminated. These include the files that terminate with .gif, .jpg, .jpeg, .css etc. This step eliminates 40018 instances. The next step of URL splitting proceeds with 24697 instances. A sample of randomly selected 2000 instances is taken for experimentation. The URL information of these instances alone is considered for further processing. Sample URLs in the dataset is shown in Figure3.

URL 1	Shuttle/countdown
URL 2	History/Apollo/
URL 3	Shuttle/Mission/sts-73/mission-hts-73.html

Figure 3. Sample URLs

The URL is then spitted into individual components. The splitting is done with the symbol slash '/' as the delimiter. The URL splits into maximum of five parts. A URL can contain one to five sub-parts. The methodology proceeds with these five parts. Feature vector formulation is carried out as the next step, the details of which are provided in the following sub-section.

C. Feature Vector Formulation

The five parts of the URL forms five fields in the feature vector. The class field describing the web page category is appended to the data. Totally, 30 categories of web pages are defined for NASA dataset in [20]. These categories are listed in Table 1.

Table1. Page Categories

Categories	Categories	Categories
Ele	Icon	shuttle/countdown
Facilities	Images	shuttle/movies
Shuttle/mission	Logistics	Software
Downs	Mdss	Statistics
base-ops	Msfc	history/Apollo
bio-med	News	history/Gemini
Facts	Pao	history/mercury
finance	Payloads	Shuttle
History	Persons	shuttle/resources
Htbin	Procurement	shuttle/technology

The class fields are appropriately added to the instances. Thus, the feature vector constitutes the URL1, URL2, URL3, URL4 and URL5 followed by the webpage category. The formulated feature vector is provided as input to the subsequent step namely classification, the details of which are presented in the following sub-section.

D. Classification

Classification [21] refers to the process of building a learning model through the labeled training samples. The test samples can then be given to the built learning model to identify its label. Various classification algorithms have been proposed in the literature for various applications. The performance of the classification algorithm is application specific. For the task in hand, C4.5 classification algorithm performs the best.

C4.5 classification algorithm [22] is a decision tree algorithm based on gain ratio evaluation metric. The procedure of C4.5 is depicted in Figure4.

```
Input:
D: Dataset comprising of N labelled training samples
A: set of attributes

Method: C4.5(D,A)
Step 1: Create a node $N$.
Step 2: If all instances in $D$ belong to the same class $C$, then
        Return $N$ as the leaf node labelled with class $C$.
Step 3: If $A$ is empty, then
        Return $N$ as a leaf node labelled with the majority class in $D$.
Step 4: For all attributes $a$ in A, calculate gain ratio as follows:
            $GainRatio(a) = \frac{Gain(a)}{SplitInfo(a)}$
            Where $splitinfo_a(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} * log_2(\frac{|D_j|}{|D|})$
Step 5:Assign $a_{best}$= attribute with maximum gain ratio
Step 6: Label node $N$ with $a_{best}$ and let it test the splitting
        criterion.
Step 7: For each outcome $j$ of the splitting criterion,
            $D_j$ = data instances in $D$ satisfying outcome $j$.
Step 8: If $D_j$ is empty, then
            Attach a leaf labelled with the majority class in $D$ to node $N$.
Step 9: Else, attach the node returned by recursively
        calling C4.5(D,A).
Step 10: return $N$.
```

Figure 4. C4.5 Procedure

C4.5 classification algorithm initiates the tree construction from the root node. The attribute that yields the highest gain ratio is assigned as the root node and the tree building continues until one of the termination criteria is met.

In the context of web page classification, many classifiers are attempted, out of which C4.5 performed the best. The performance evaluation of the classifiers is presented in the following sub-section.

E. Performance Evaluation

Evaluation of the classification algorithms and hence the learning model built is dome through k-fold cross validation [23]. Cross validation technique divides the data into k folds. During the first iteration, the instances of fold 1 to k-1 are used for training while the instances from fold k are given for testing. During the second iteration, training samples from fold 2 to k is provided for training whereas the samples from fold 1 are utilized for testing and the procedure repeats for k times. The accuracy is computed over all the iterations.

Accuracy is defined as the ratio of total number correct predictions to the total number of predictions. It is given by Equation 1.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ Number\ of\ predictions} \qquad (1)$$

The performance of the classifiers towards web page categorization is assessed through 3 fold cross validation.

The proposed approach exhibits a justifiable performance in the view of web page classification. The relevant results are reported in the following section.

IV. RESULTS AND DISCUSSION

The experiments are carried out in Tanagra, an open source data mining tool [24] [25]. Various classification algorithms have been tested in the context of web page categorization. They include C4.5 [22], Random Tree [26], Rule Induction [27], ID3 [28], C-RT, Naive Bayes [29], Decision List and CS-CRT. These classifiers are applied on the feature vector formed from the pre-processed data. The performance is evaluated through accuracy %. The 3 fold cross validation is used for the assessment. Table 2 reports the accuracy % of all the classifiers considered

Table 2. Performance Evaluation of Classifiers (Accuracy %)

Classifier	**Accuracy (%)**
C4.5	**99.80**
Random Tree	97.00
Rule Induction	97.00
ID3	97.00
C-RT	76.50
Naive Bayes	96.50
Decision List	73.70
CS-CRT	76.50

Table 2 shows that C4.5 performs superior to the other classification algorithm is this context yielding an accuracy of 99.80%.

V. CONCLUSION

Web usage mining reveals interesting facts on mining the usage patterns of the users. Web page categorization helps in apt advertisements and recommendations to the user viewing the web page. Web page categorization is attempted through URL information through the proposed approach. The approach comprises of data pre-processing incorporating removal of irrelevant data and URL splitting, feature vector formulation, classification, performance evaluation and prediction of class label of the web page. Experimentation has been performed on the publicly available NASA log through various classifiers. C4.5 yields the maximum accuracy of 99.8% on 3 fold cross validation when categorizing the web pages into 30 distinct classes. The high performance justifies the efficiency of the proposed approach.

VI. REFERENCES

[1]. J. Han and M. Kamber. 2011. Data Mining – Concepts and Techniques, Morgan Kauffmann Publishers, 3rd Edition.

[2]. R. Geetharamani, P. Revathy and S.G. Jacob. 2015. Prediction of Users Webpage Access Behaviour Using Association Rule Mining. Sadhana, 40(8), pp.2353-2365.

[3]. R. Geetha Ramani and B. Lakshmi. 2013. Multi-Class Classification for Prediction of Retinal Diseases (Retinopathy and Occlusion) from Fundus Images. In Proceedings of ICKM, 13, pp. 122-134.

[4]. P. Nancy, R. G. Ramani and S.G. Jacob. 2011. Discovery of Gender Classification Rules for Social Network Data using Data Mining Algorithms. In Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC’2011).

[5]. C. Robert, B. Mobasher and J. Srivastava. 1997. Web Mining: Information and Pattern Discovery on the World Wide Web. in Proceedings of 9th International Conference on Tools with Artificial Intelligence.

[6]. 6R. Kosala and H. Blockeel. 2000. Web Mining Research: A Survey. ACM Sigkdd Explorations Newsletter, 2(1), pp. 1-15.

[7]. J. Srivastava et al. 2000. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. ACM Sigkdd Explorations Newsletter, 1(2) , pp. 12-23.

[8]. J. Faustina and Santosh Kumar Gupta. 2012. Web Content Mining Techniques: A Survey. International Journal of Computer Applications, 47(11) .

[9]. D. Costa, M. Gomes and Z. Gong. 2005. Web Structure Mining: An Introduction. IEEE International Conference on Information Acquisition.

[10]. G. Kesavaraj and S. Sukumaran. 2013. A study on classification Techniques in Data Mining. In Proceedings of 2013 4th IEEE International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1-7.

[11]. P. Nancy and R.G. Ramani. 2012. Discovery of Patterns and Evaluation of Clustering Algorithms in Social Network Data (Face book 100 universities) through Data Mining Techniques and Methods. International Journal of Data Mining & Knowledge Management Process, 2(5), p.71.

[12]. J. Hipp, U. Güntzer and G. Nakhaeizadeh. 2000. Algorithms for Association Rule Mining—A General Survey and Comparison. ACM sigkdd explorations newsletter, 2(1), pp. 58-64.

[13]. J. He, A.H. Tan and C.L. Tan. 2000. Machine Learning Methods for Chinese Web Page Categorization. In Proceedings of the second workshop on Chinese language processing: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 12, pp. 93-100. Association for Computational Linguistics.

[14]. S. Noh, H. Seo, J. Choi, K. Choi and G. Jung. 2003. Classifying Web Pages using Adaptive Ontology. In IEEE International Conference on Systems, Man and Cybernetics, vol. 3, pp. 2144-2149.

[15]. S.M. Kamruzzaman. 2006. Web Page Categorization using Artificial Neural Networks. In Proceedings of the 4th international conference on Electrical Engineering and 2nd Annual Paper Meet, pp. 96-99. arXiv preprint arXiv:1009.4991.

[16]. J.A. Mangai and V.S Kumar. 2011. A Novel Approach for Web Page Classification using Optimum Features. IJCSNS, 11(5), p.252.

[17]. S. Kavitha and M.S. Vijaya. 2013. Web Page Categorization using Multilayer Perceptron with Reduced Features. International Journal of Computer Applications, 65(1).

[18]. M.F. Arlitt and C.L. Williamson. 1996. Web Server Workload Characterization: The Search for Invariants. ACM SIGMETRICS Performance Evaluation Review, 24(1), pp.126-137.

[19]. NASA-HTTP. Available online at http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html

[20]. G. Poornalatha and P. S. Raghavendra. 2012. Web Page Prediction by Clustering and Integrated Distance Measure. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). IEEE Computer Society.

[21]. R.G. Ramani, B. Lakshmi and S.G. Jacob. 2012. Automatic Prediction of Diabetic Retinopathy and Glaucoma through Retinal Image Analysis and Data Mining Techniques. In Proceedings of 2012 IEEE International Conference on Machine Vision and Image Processing (MVIP), pp. 149-152.

[22]. J.R. Quinlan. 2014. C4. 5: Programs for Machine Learning. Elsevier.

[23]. R.G. Ramani, B. Lakshmi and S.G. Jacob. 2012. Data Mining Method of Evaluating Classifier Prediction accuracy in Retinal Data. In Proceedings of the 2012 IEEE International Conference on Computational Intelligence & Computing Research (ICCIC), pp. 1-4.

[24]. R Rakotomalala. 2005. TANAGRA: A Free Software for Rresearch and Academic Purposes. in Proceedings of EGC'2005, RNTI-E-3, 2, pp.697-702. (in French)

[25]. Tanagra. Available online at https://eric.univ-lyon2.fr/ricco/tanagra/en/tanagra.html

[26]. L. Breiman. 2001. Random Forests. Machine Learning, 45(1), pp.5-32.

[27]. W.W. Cohen. 1995. Fast Effective Rule Induction. In Proceedings of the twelfth international conference on machine learning, pp. 115-123.

[28]. X. Wu et al. 2008. Top 10 Algorithms in Data Mining. Knowledge and information systems, 14(1), pp.1-37.

[29]. K.P. Murphy. 2006. Naive Bayes Classifiers. University of British Columbia.