International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.3, No.6, June 2014 DOI:10.15693/ijaist/2014.v3i6.130-136

Effective Evaluation of Classification of Indigenous News Using Decision Table and OneR Algorithm

S. R. Kalmegh
Associate Professor, P.G. Department of Computer Science, SGBAU, Amravati (M.S.), India.

S. N. Deshmukh
Associate Professor, Department of Computer Engineering & I.T., DBAMU, Aurangabad (M.S.), India.

Abstract: Classification may refer to categorization, the process in which ideas and objects are recognized, differentiated, and understood. An algorithm that implements classification, especially in a concrete implementation, is known as a **classifier**. Classification is an important data mining technique with broad applications. It classifies data of various kinds. This paper has been carried out to make a performance evaluation of Decision Table and OneR classification algorithm. The paper sets out to make comparative evaluation of classifiers Decision Table and OneR in the context of dataset of Indian news to maximize true positive rate and minimize false positive rate. For processing Weka API were Used. The results in the paper on dataset of news also show that the efficiency and accuracy of OneR is good than Decision Table.

Keywords: dataset, dynamic multimedia content, e-Learning, Decision Table, OneR

1.INTRODUCTION

As the latest stage of learning and training evolution, e-Learning is supposed to provide intelligent functionalities not only in processing multi-media education resources but also in supporting context-sensitive pedagogical education processes.

In recent years, people have been used to using In this case, there is a need of a design of a framework which can integrate dynamic multimedia content to the existing e-contents. This paper discusses the methodology for such integration. In order to get the details of this methodology, this paper is organized into five parts. First part discusses the concept of e-learning followed by the literature required for analysis of methods implemented. Fourth one discusses the technique of classification. Fifth one is System Design followed by datasets used for analysis. Seventh is the Performance Analysis and then conclusions.

2. E-LEARNING

E-learning is a new education concept by using the Internet technology, it deliveries the digital content, provides a learner-orient environment for the teachers and students. The e-learning promotes the construction of life-long learning opinions and learning society.

It means:

the Internet as an important information channel for working and living. More and more daily activities are relying on the global network than before, for example, e-Business, e-Government, e-Science, and e-Learning. Among these e-Activities, e-Learning has been regarded as a fast growing research and application area with huge market potential. However, e-Learning is different from other e-Activities for its involvement of precise information retrieval, systematic knowledge management, and pedagogical process. These features make e-Learning systems more complicated than basic web-based information systems, which consequently need integrated solutions to address those issues together, especially when multimedia education resources are more and more popular. As people have experienced on the Internet, finding the right information is not an easy thing, and finding multimedia resources which are semantically relevant to requests is even harder. The limitation of HTML in information representation is an essential issue, since HTML was designed to represent human readable literal information rather than carrying machine readable semantic information of literal and multimedia resources. In a practical e-Learning scenario, the information and knowledge exchange is more frequent than that in a normal information retrieval case on the Web, because people just naturally treat an e-Learning system as more organized information and knowledge base rather than a massive global network.

1. E-learning is a new education concept; it may differ from the old educational concept.
2. Delivery of the digital content is the main characters of e-learning.
3. This definition extends the environment on the Internet. It means that the Internet provides a learning environment for the students and teachers. This environment is learner-oriented, so we can throw out the thoughts of traditionally teacher-center's instruction in classroom.
4. As a new concept of education, e-learning gives a condition for us to realize the life-long learning principle and help us to build a more real learning society.

E-learning plays a major role in high education for the reason of fast need of high education.

3. LITERATURE SURVEY

There are a number of e-Learning software systems on the market such as WebCT Blackboard, Learning Space, and PageOut. The most common

function offered by those systems is courseware management, which is basically file-level content management. Although some of those systems (e.g., WebCT) claim to be able to integrate with certain academic information systems, the underlying computing technology is still at superficial level.

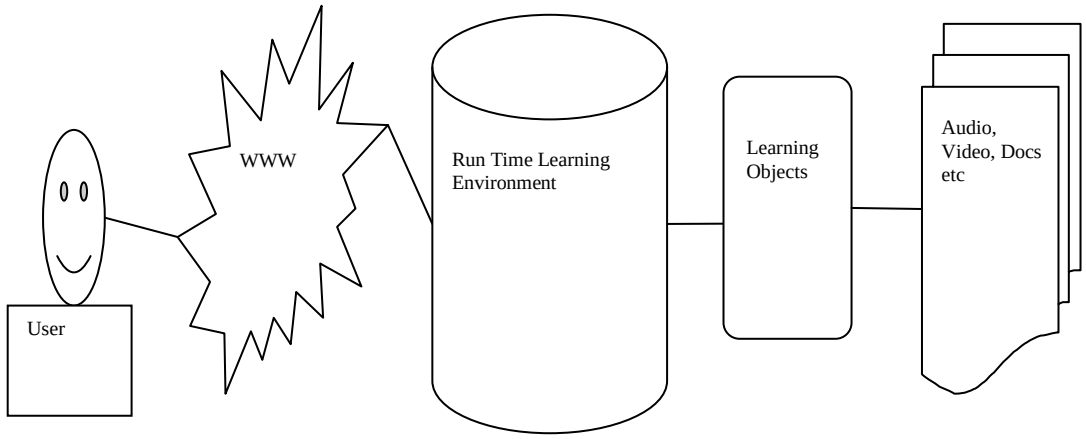


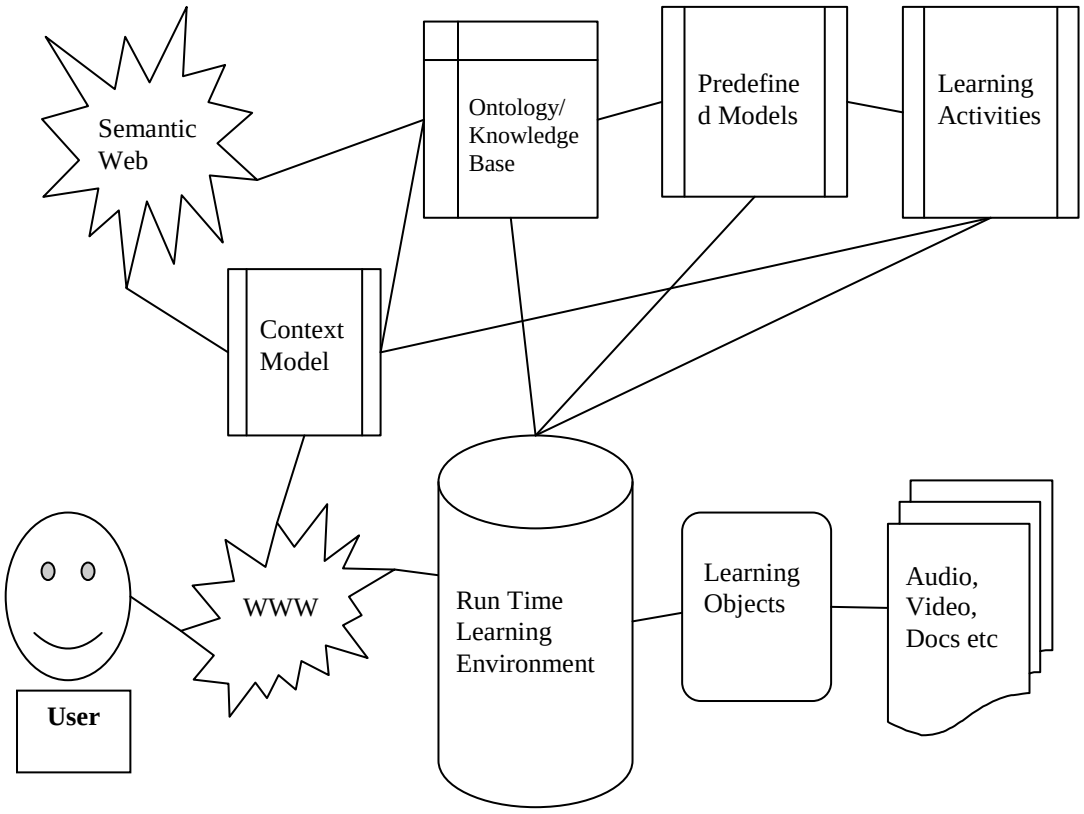
Fig1: Traditional e-Learning System

The major implementation that includes the intelligence in e-Learning is ConKMEL. To resolve the knowledge integration and management problem in multimedia e-Learning, it has proposed a semantic context aware approach, which features an integrated contextual knowledge management framework to support intelligent e-Learning.[1]

Traditional web-based e-learning systems use a web browser as the interface. Through run-time learning environments (either compatible or incompatible with SCORM) [2],[3], users could access the learning objects, which are directly linked to multimedia learning resources such as lecture video/audio, presentation slides and reference documents. A flow in traditional e-Learning system is given in Fig 1.

Weihong Huang et. al. has proposed an intelligent semantic e-Learning framework which presents semantic information processing, learning process support and personalized learning support

Weihong Huang et. al. has proposed an intelli **Fig** gent semantic e-Learning framework which presents semantic information processing, learning process support and personalized learning support issues in an integrated environment.



Architecture of the above framework is as given below in Fig 2.

Fig 2: Semantic e-Learning Framework

In addition to the traditional learning information flow, three new components namely semantic context model, intelligent personal agents and conceptual learning theories are introduced to bring in more intelligence Intelligent personal agents perform adequate personal trait information profiling and deliver personalised learning services. Semantic context model uses semantic information for static resource and dynamic process retrieves information from WWW and the future Semantic Web, referring to ontologies or knowledge bases. [4]

4.CLASSIFICATION

Classification may refer to categorization, the process in which ideas and objects are recognized, differentiated, and understood. An algorithm that implements classification, especially in a concrete implementation, is known as a **classifier**. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category.

In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as *clustering* or cluster analysis, and involves grouping data into categories based on some measure of inherent similarity (e.g. the distance between instances, considered as vectors in a multi-dimensional vector space).

Classification (also known as classification trees or decision trees) is a data mining algorithm that creates a step-by-step guide for how to determine the output of a new data instance. The tree it creates is exactly that: a tree whereby each node in the tree represents a spot where a decision must be made based on the input, and to move to the next node and the next until one reach a leaf that tells the predicted output. Sounds confusing, but it's really quite straightforward. Let's look at an example.

3.1 Decision Table Classifiers :

The entire problem of learning decision tables consists of selecting the right attributes to be included. Usually this is done by measuring the table's

crossvalidation performance for different subsets of attributes and choosing the bestperforming subset. Fortunately, leave-one-out cross-validation is very cheap for this kind of classifier. Obtaining the cross-validation error from a decision table derived from the training data is just a matter of manipulating the class counts associated with each of the table's entries, because the table's structure doesn't change when instances are added or deleted. The attribute space is generally searched by best-first search because this strategy is less likely to get stuck in a local maximum than others, such as forward selection. [5]

Decision Table are one of the simplest hypothesis spaces possible and usually they are easy to understand. Decision Table builds a decision table majority classifier. It evaluates feature subsets using best-first search and can use cross-validation for evaluation. An option uses the nearest-neighbor method to determine the class for each instance that is not covered by a decision table entry, instead of the table's global majority, based on the same set of features. [6] [7]

3.2 OneR Classifiers :

OneR is a simple and a very effective classification algorithm frequently used in machine learning applications. Even though OneR is difficult to be improved further due to its simplicity, it can be enhanced by providing better methods for handling some of the exceptions. There is an easy way to find very simple classification rules from a set of instances, called 1R for 1-rule, due to its high accuracy. Perhaps this is because the structure underlying many real-world datasets is quite rudimentary, and just one attribute is sufficient to determine the class of an instance quite accurately. In any event, it is always a good plan to try the simplest things first. Generates a one-level decision tree expressed in the form of a set of rules that all test one particular attribute. 1R is a simple, cheap method that often comes up with quite good rules for characterizing the structure in data. It turns out that simple rules frequently achieve surprisingly.

The idea is this: We make rules that test a single attribute and branch accordingly. Each branch corresponds to a different value of the attribute. It is obvious what the best classification to give each branch is, Use the class that occurs most often in the training data. Then the error rate of the rules can easily be determined. Just count the errors that occur on the training data that is, the number of instances that do not have the majority class. Each attribute generates a different set of rules, one rule for every value of the attribute. [5] [7] [8]

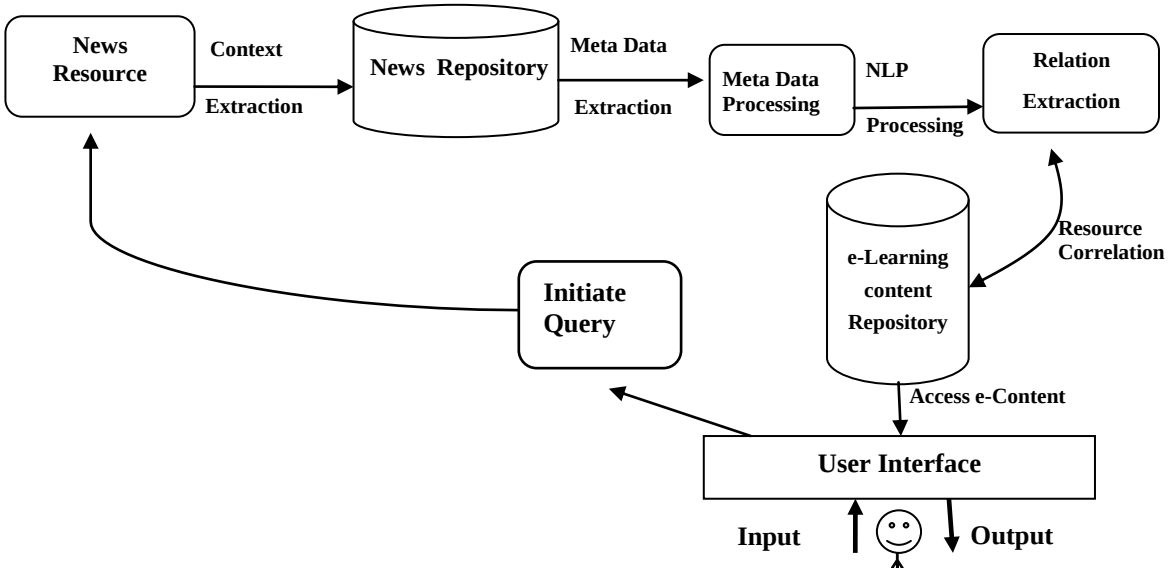


Fig 3. Flow diagram of the model

5. SYSTEM DESIGN

We designed a model based on the machine learning and XML search. In order to

co-relate News with the categories a model based on the machine learning and XML search was designed. Flow diagram of the model for news resources is shown below.

As a input to the model, various news resources are considered which are available online like the news in Google news repository or online paper like Times of India, Hindustan Times etc. Around 649 news were collected on above repository. In order to extract context processing in the above flow diagram. Title of the also contains useful information in the abstract form, The title also can be considered as Metadata. The title of the news is processed using NLP libraries (Standford NLP Library) to extract various constituents of it. The output of NLP process was also used to co-relate the News (textual, audio, video) to the concern e-learning contents. This process can be initiated automatically when the user access any content from e-Learning data repository.

As shown in the figure, a news resource is processed to correlate with the e-Contents available. On the similar way, other text resources can be added directly with the e-Content in the repository, Image or Video resource can be processed for meta-data available. And thus can be searched with the related e-Contents.

6. DATA COLLECTION

Hence it was proposed to generate indigenous data. Hence the national resources were used for the research purpose. Data for the purpose of research has been collected from the various news which are available in various national and regional newspapers available on internet. They are downloaded and after reading the news they are manually classified into 7 (seven) categories. There were 649 news in total. The details are as shown in Table 1.

News Category	Actual No. Of News
Business	123
Criminal	82
Education	59
Medical	46
Politics	153
Sports	147
Technology	39
Total	**649**

from the news and co-relate it with the proper e-content, the News was process with stemming and tokenization on the news contents. The news then was converted into the term frequency matrix for further analysis purpose. Based on this data, features (i.e. metadata) were extracted so that contextual assignment of the news to the appropriate content can be done. This process is known as metadata

The attributes consider for this classification is the topic to which news are related; the statements made by different persons; the invention in Business, Education, Medical, Technology; the various trends in Business; various criminal acts e.g. IPC and Sports analysis. During classification some news cannot be classified easily e.g.

1. Political leader arrested under some IPC code,
2. Some invention made in medicine and launched in the market & business done per annum.

7. PERFORMANCE ANALYSIS

The News so collected needed a processing. Hence as given in the design phase, all the news were processed for stop word removal, stemming, tokenization and ultimately generated the frequency matrix. Stemming is used as many times when news is printed, for a same there can be many variants depending on the tense used or whether it is singular or plural. Such words when processed for stemming, generates a unique word. Stop words needs to be removed as they do not contribute much in the decision making process. Frequency matrix thus generated can be processed for generating a model and the model so generated was used in further decision process.

With the model discussed above, two classifier J48 and Ridor were used on the data set of 649 news. For processing Weka APIs were used. The result after processing is given in the form of confusion matrix which is shown in Table 1. and Table 2.

Table 1. Confusion Matrix using Decision Table classifier

Classified as →	Education	Business	Criminal	Technology	Politics	Medical	Sports
Education	44	2	3	0	4	0	6
Business	4	79	2	0	9	0	29
Criminal	11	2	55	0	8	1	5
Technology	5	9	2	9	2	0	12
Politics	43	0	3	0	105	0	2
Medical	6	2	0	0	3	31	4
Sports	2	2	1	0	0	1	141

Table 2. Confusion Matrix using OneR classifier

Classified as →	Education	Business	Criminal	Technology	Politics	Medical	Sports
Education	59	0	0	0	0	0	0
Business	0	123	0	0	0	0	0
Criminal	0	0	82	0	0	0	0
Technology	0	0	0	39	0	0	0
Politics	0	0	0	0	153	0	0
Medical	0	0	0	0	0	46	0
Sports	0	0	0	0	0	0	147

Comparative analysis of the confusion matrix table shown above is given Table 3. below. It clearly indicates that OneR the best suite for such type of data.

Table 3. Showing correct and wrong Prediction of Classifier.

Classifier →		J48		Ridor	
News Category	**Actual No. Of News**	**Correct**	**Wrong**	**Correct**	**Wrong**
Business	123	122	01	104	19
Criminal	82	74	08	67	15
Education	59	53	06	38	21
Medical	46	42	04	28	18
Politics	153	143	10	134	19
Sports	147	143	04	136	11
Technology	39	31	08	13	26
Total	649	608	41	520	129
Percentage →		**93.68%**	**6.32%**	**80.12%**	**19.88%**

8. CONCLUSION

As shown in previous discussion identification of news from dynamic resources can be done with the propose model we use two classifier i.e. Decision Table and OneR to analyze the data sets. As a result it is found that OneR algorithm perform well in categorizing in the all the News. Ridor algorithm perform well in categorizing in the news related to Education, Medical and Sports. For overall data set detection rate for OneR is 100% and whereas Decision Table is 71.49%. Hence OneR is good classifier as compare to Decision Table classifier.

REFERENCES

[1] Weihong Huang, Alain Mille, ConKMel: a contextual knowledge management framework to support multimedia e-Learning, Published online: 8 July 2006, Springer Science + Business Media, LLC 2006

[2] SCORM (2003) Advanced distributed learning initiative, sharable content object reference model (SCORM). http://www.adlnet.orgl/

[3] LOM (1999) IEEE P1484.12 learning object metadata working group, learning object metadata. http://ltsc.ieee.org/wg12/

[4] Weihong Huang, David Webster, Dawn Wood and Tanko Ishaya, "*An intelligent semantic e-learning framework using context-aware Semantic Web technologies", British Journal of Educational Technology, Vol 37 No 3, pp* 351–373, 2006

[5] Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining Practical Machine Learning Tools and Techniques, Third Edition, Morgan Kaufmann Publishers is an imprint of Elsevier

[6] Ron Kohavi. The power of Decision Tables. European Conference on Machine Learning (ECML) 1995.

[7] M. Thangaraj, C.R.Vijayalakshmi. Performance Study on Rule-based Classification Techniques across Multiple Database Relations. International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 5– No.4, March 2013

[8] Gaya Buddhinath ,Damien Derry. A Simple Enhancement to One Rule Classification Department of Computer Science & Software Engineering University of Melbourne