**International Refereed Journal of Engineering and Science (IRJES)**
*ISSN (Online) 2319-183X, (Print) 2319-1821*
*Volume 3, Issue 6 (June 2014), PP.79-84*

# Categorical Identification of Indian News Using J48 and Ridor Algorithm

S. R. Kalmegh[1], S. N. Deshmukh[2]

[1]*Associate Professor, P.G. Department of Computer Science, SGBAU, Amravati (M.S.), India.*
[2]*Associate Professor, Department of Computer Engineering & I.T., DBAMU, Aurangabad (M.S.), India.*

**Abstract:-** 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 J48 and Ridor classification algorithm. The paper sets out to make comparative evaluation of classifiers J48 and Ridor 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 J48 and Ridor is good.

**Keywords:-** dataset, dynamic multimedia content, e-Learning, J48, Ridor

## I. 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 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.

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 four parts. First part discusses the literature required for analysis of methods implemented. Second one discusses the datasets used for analysis which is followed by Performance Analysis and then conclusions.

## II. 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.

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.

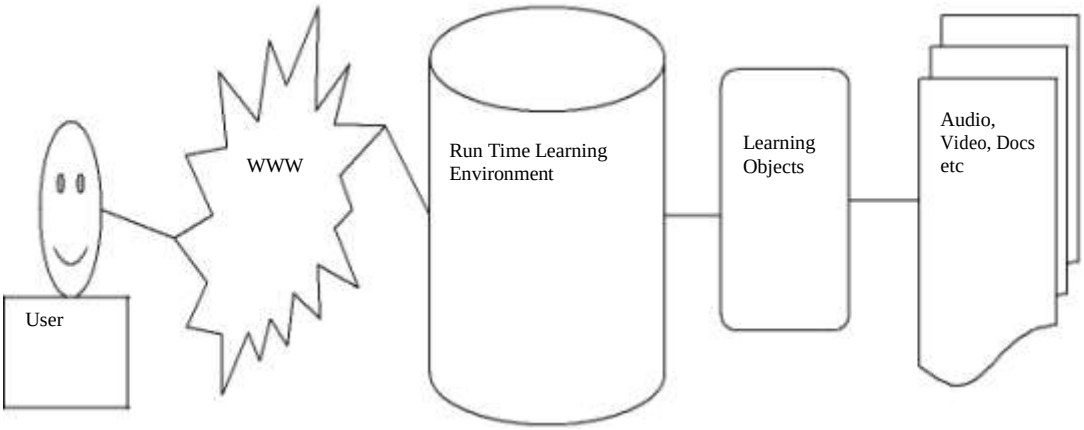


**Fig 1: Traditional e-Learning System**

Weihong Huang et. al. has proposed an intelligent 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. 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]

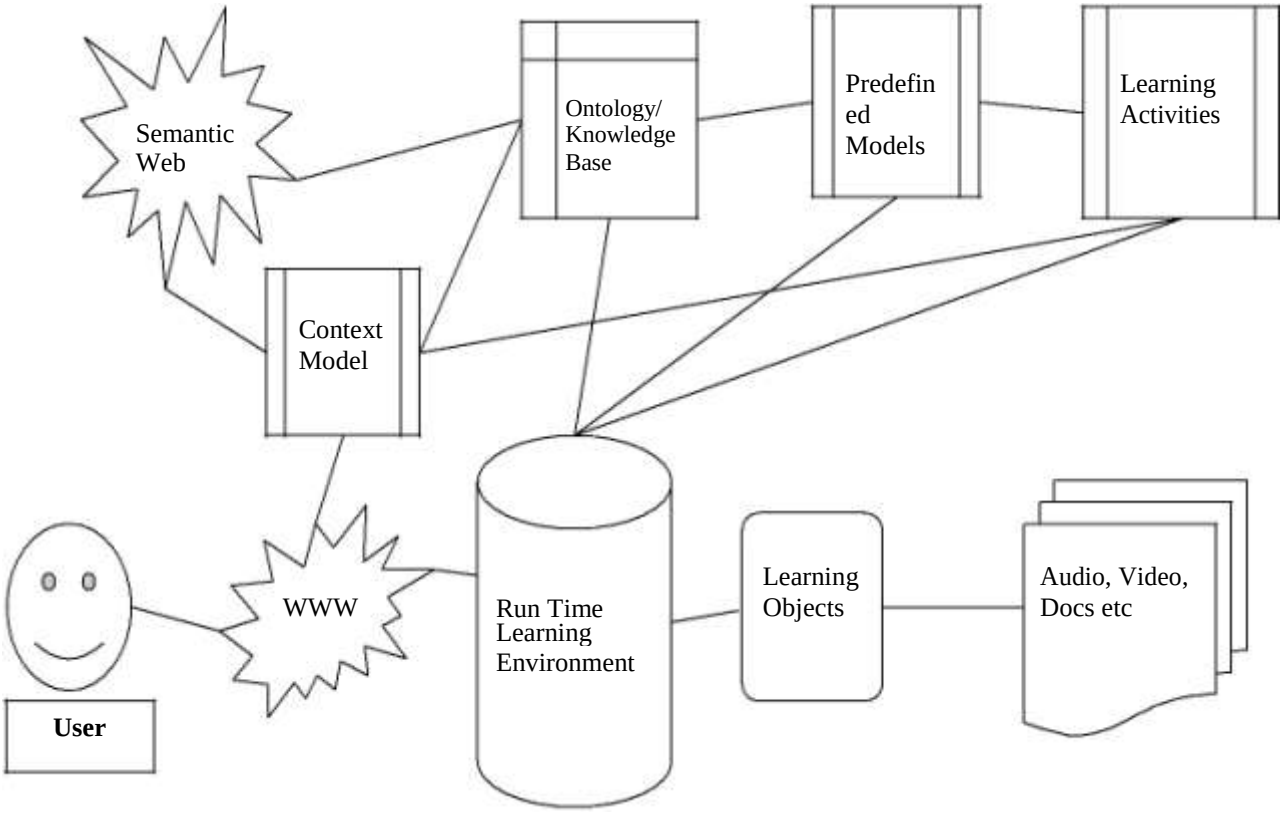


**Fig 2: Semantic e-Learning Framework**

## III. 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 Ridor Classifiers:*

Ripple Down Rule learner (RIDOR) is also a direct classification method. Ridor learns rules with exceptions by generating the default rule, using incremental reduced-error pruning to find exceptions with the smallest error rate, finding the best exceptions for each exception, and iterating. It generates a default rule first and then the exceptions for the default rule with the least (weighted) error rate. Then it generates the "best" exceptions for each exception and iterates until pure. Thus it performs a tree-like expansion of exceptions. The exceptions are a set of rules that predict classes other than the default. IREP is used to generate the exceptions. Incremental Reduced Error Pruning IREP is used to create the exceptions. [5] [6] [7]

### *3.2 J48 Classifiers:*

J4.8 actually implements a later and slightly improved version called C4.5 revision 8, which was the last public version of this family of algorithms before the commercial implementation C5.0 was released.

The J48 Decision tree classifier follows the following simple algorithm. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then we terminate that branch and assign to it the target value that we have obtained.

A flow-chart-like tree structure internal node denotes a test on an attribute Branch represents an outcome of the test Leaf nodes represent class labels or class distribution. Decision tree generation consists of two phases Tree construction At start, all the training examples are at the root. Partition examples recursively based on selected attributes Tree pruning Identify and remove branches that reflect noise or outliers. Use of decision tree: Classifying an unknown sample Test the attribute values of the sample against the decision tree. [5] [8]

## IV. 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.

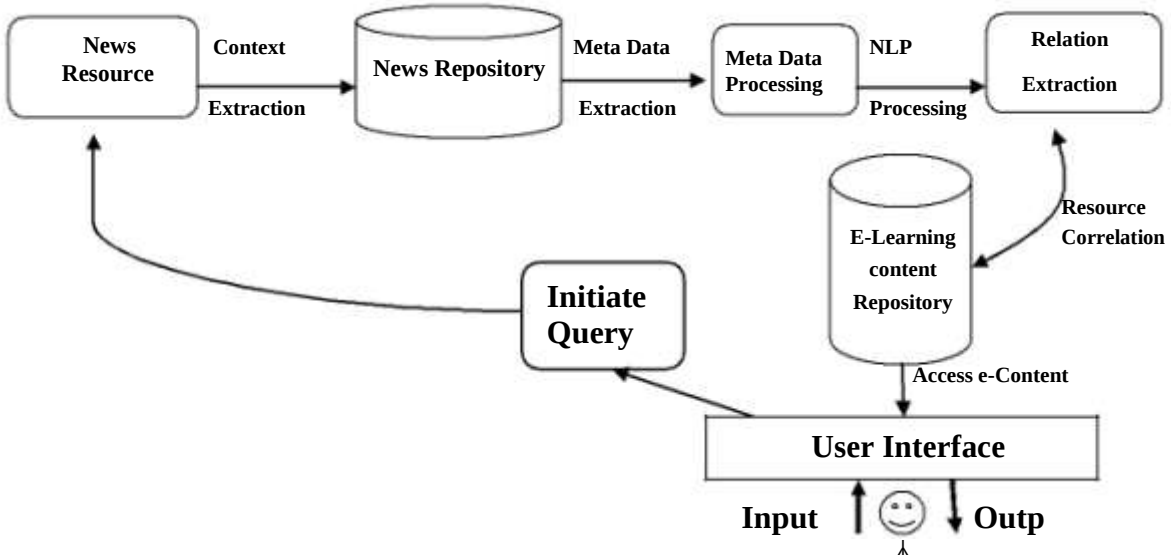


**Fig 3. Flow diagram of the model**

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 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 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.

## V. 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.

**Table 1: Showing the Category of the News**

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

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.

## VI. 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 J48 classifier**

| CLASSIFIED AS → | a | b | c | d | e | f | g |
|---|---|---|---|---|---|---|---|
| **a = Education** | 53 | 1 | 1 | 2 | 1 | 0 | 1 |
| **b = Business** | 0 | 122 | 0 | 0 | 1 | 0 | 0 |
| **c = Criminal** | 1 | 2 | 74 | 3 | 1 | 0 | 1 |
| **d = Technology** | 4 | 0 | 1 | 31 | 1 | 1 | 1 |
| **e = Politics** | 3 | 0 | 4 | 2 | 143 | 1 | 0 |
| **f = Medical** | 0 | 1 | 1 | 2 | 0 | 42 | 0 |
| **g = Sports** | 0 | 0 | 0 | 3 | 1 | 0 | 143 |

**Table 2. Confusion Matrix using Ridor classifier :**

| CLASSIFIED AS → | a | b | c | d | e | f | g |
|---|---|---|---|---|---|---|---|
| **a = Education** | 38 | 6 | 3 | 0 | 6 | 0 | 6 |
| **b = Business** | 1 | 104 | 1 | 2 | 13 | 0 | 2 |
| **c = Criminal** | 3 | 5 | 67 | 0 | 4 | 3 | 0 |
| **d = Technology** | 1 | 11 | 0 | 13 | 7 | 0 | 7 |
| **e = Politics** | 2 | 4 | 10 | 0 | 134 | 1 | 2 |
| **f = Medical** | 1 | 11 | 1 | 0 | 1 | 28 | 4 |
| **g = Sports** | 1 | 9 | 1 | 0 | 0 | 0 | 136 |

Comparative analysis of the confusion matrix table shown above is given Table 3. below. It clearly indicates that J48 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%** |

## VII. 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. J48 and Ridor to analyze the data sets. As a result it is found that J48

algorithms perform well in categorizing in the news related to Business and Sports. Ridor algorithm perform well in categorizing in the news related to Business, Politics and Sports. For overall data set detection rate for J48 is 93.68% and whereas Ridor is 80.12%. Hence J48 is good classifier as compare to Ridor 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]. C. Lakshmi Devasena, T. Sumathi, V.V. Gomathi and M. Hemalatha. Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set. Bonfring International Journal of Man Machine Interface, Vol. 1, Special Issue, December 2011

[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]. Anshul Goyal and Rajni Mehta. Performance Comparison of Naïve Bayes and J48 Classification Algorithms. International Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7 No.11 (2012)