International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 4, Issue 3, May-June 2015 ISSN 2278-6856

Metadata Based Classification and Analysis of Large Scale Web Videos

[1]Siddu P. Algur , *Prashant Bhat[2]

[1]Department of Computer Science, Rani Channamma University,
Belagavi-591156, Karnataka, India

[2] Department of Computer Science, Rani Channamma University,
Belagavi-591156, Karnataka, India

Abstract

The astonishing growth of videos on the Internet such as YouTube, Yahoo Screen, Face Book etc, organizing videos into categories is of paramount importance for improving user experience and website utilization. In this information age, video information is the rapidly sharing by the people through social media websites such as YouTube, Face Book, yahoo Screen etc. Different categories of web video are shared on social websites and used by the billions of users all over the world. The classification/partitioning of web videos in terms of length of the video, ratings, age of the video, number of comments etc, and analysis of this web video as a unstructured complex data is a challenging task. In this work we propose effective classification model to classify each category of web-videos (Ex- 'Entertainment', 'People and Blogs', 'Sports', 'News and Politics', 'Science and Technology' etc) based on other web metadata attributes as splitting criteria. An attempt is made to extract metadata from web videos. Based on the extracted metadata, web videos are classified/partitioned into different categories by applying data mining classification algorithms such as and Random Tree and J48 classification model. The classification results are compared and analyzed using cost/benefit analysis. Also the results demonstrate classification of web videos depends largely on available metadata and accuracy of the classification model. Classification/partitioning of web-based videos are important task with many applications in video search and information retrieval process. However, collecting metadata required for classification model may be prohibitively expensive. The experimental difficulties arise from large data diversity within a category is pitiable of metadata and dreadful conditions of web video metadata.

Keywords: Web Video Mining, Web Video Metadata, Random Tree Algorithm, J48 Algorithm, Web Video Categories.

1. INTRODUCTION

The incredible rapid growth of videos on the Internet such as YouTube, Yahoo Screen, and Face Book etc organizing videos into different classes is a dominant significance for improving user experience (searching, access, upload, online watch etc) and website monetization nowadays. Classification of web videos is an increasingly outstanding area of research, growing with the quantity of videos shared through some social sites such as YouTube, Yahoo Screen etc. As much as its importance, web based video classification poses serious challenges to computer vision researchers [1]. The difficulties are multifold, including large data diversity within a class/category [1], insufficient metadata, and dreadful conditions of quality in some web videos. Some related works have been implemented to classify web videos based on image and signal processing techniques. Also it is observed that, a very less number of works have been attempted to classify web videos based on metadata. This is because insufficient metadata available for the classification of web videos. Some of the web videos don't have sufficient metadata, because the uploader of the video ignores to give sufficient metadata while uploading video to the websites. This will lead to serious problem for users. The authors Siddu P. Algur, Prashant Bhat and Suraj Jain [2], have proposed a model to construct metadata for web videos and suggested the uploaders to follow metadata construction model/guidelines while uploading videos to the web. This will helps user to search, retrieve web videos effectively and accurately, and to know about the videos without watching it. The classification problem consists of four major components [3]. The first component is a categorical outcome or "dependent" variable. This variable is the characteristic which we hope to predict, based on the "independent" or "predictor" variables. The classic outcome variables are Entertainment, Sports, News and Politics, People and Blogs, Music, Comedy, etc. There are 16 such categories found during the collection of dataset. The second component of classification problem is the "predictor" or "independent" variables. These are the characteristics which are potentially related to the outcome variable of interest. In general, there are many possible predictor variables. In this classification experiment, the second components are web metadata such as number of comments, likes, ratings, average rate etc. The third component of the classification problem is the learning dataset. This is a dataset which includes values for both the outcome and predictor variables. The fourth component of the classification problem is the test or future dataset, which consists of web metadata datasets for which we able to make accurate predictions.

Many classification models/algorithms and data mining and machine learning tools are developed in recent years. In this work, the web video metadata are extracted and classified/partitioned based on available metadata using

Random Tree (RT) and J48 classification algorithms. The classification results are compared and cost/benefit of each web video category is analyzed.

The rest of the paper is organized as follows: The section 2 represents related works on the classification of web videos, section 3 represents proposed web video classification methodology, section 4 represents performance evaluation analysis of classification models and cost/benefit analysis of each category of web videos, and finally section 5 represents conclusion and future enhancements.

2. RELATED WORKS

This section represents some related previous works which are been implemented to classify web videos using metadata. The authors Yang Song, Ming Zhao, Jay Yagnik, and Xiaoyun Wu [1], worked on large scale video taxonomic classification system, which utilizes the category taxonomic structure in training and in interpreting the classification results and a novel scheme is proposed to adapt the web-classifiers to video domain. Video content based features are integrated with text-based features to gain power in the case of degradation of one type of features. Evaluation on videos from hundreds of categories shows that the proposed algorithms generate significant performance improvement over text classifiers or classifiers trained using only video content based features.

The authors Siddu P. Algur, Prashant Bhat and Suraj Jain [2], described significance of web video descriptive metadata, presented an effective and efficient method for construction and extraction of web video descriptive metadata. The proposed method demonstrated the effectiveness of constructing the descriptive metadata with timeline for a domain specific web video. The papers also suggested the construction of event specific and object specific metadata and which are considered to be very effective. Using proposed descriptive metadata model, users may process the video contents effectively and efficiently.

The authors John R. Zhang, Yang Song and Thomas Leung [4], explored an approach which exploits YouTube video co-watch data to improve the performance of a video taxonomic classification system. A graph is built whereby edges are created based on video co-watch relationships and weakly-labeled videos are selected for classifier training through local graph clustering. Evaluation is performed by comparing against classifiers trained using manually labeled web documents and videos. We find that data collected through the proposed approach can be used to train competitive classifiers versus the state of the art, particularly in the absence of expensive manually-labeled data.

The authors Chunneng Huang, Tianjun Fu and Hsinchun Chen [5], proposed a text-based framework for video content classification of online-video sharing websites. Different types of user-generated data (e.g., titles, descriptions, and comments) were used as proxies for online videos, and three types of text features (lexical, syntactic, and content-specific features) were extracted. Three feature-based classification techniques (C4.5, Naïve Bayes, and SVM) were used to classify videos. To evaluate the proposed framework, user-generated data from candidate videos, which were identified by searching user-given keywords on YouTube, were first collected. Then, a subset of the collected data was randomly selected and manually tagged by users as our experiment data. The experimental results showed that the proposed approach was able to classify online-videos based on users' interests with accuracy rates up to 87.2%, and all three types of text features contributed to discriminating videos. SVM outperformed C4.5 and Naïve Bayes in their experiments.

Automatic categorization of videos in a Web-scale unconstrained collection such as YouTube is a challenging task. A key issue is how to build an effective training set in the presence of missing, sparse or noisy labels. In this regard, the authors Zheshen Wang, Ming Zhao, Yang Song, Sanjiv Kumar, and Baoxin Li [5], proposed to achieve this by first manually creating a small labeled set and then extending it using additional sources such as related videos, searched videos, and text-based web pages.

3. METHODOLOGY

The online video sharing websites like YouTube, Face Book, and Yahoo Screen etc are sources of huge number of all category videos. The web metadata of online videos are extracted using Info Extractor tool. This metadata includes uploader information, category, comments, ratings, length of the video, descriptions about content of the video etc. The descriptions, category (constant metadata attributes) and uploader information are uploader dependent metadata which are given by the uploader of the video at the time of uploading. Remaining web metadata attribute variables such as comments, ratings, likes, dislikes etc are explicit metadata which are independent from the uploader. The uploader independent metadata variables change their values with the time, whereas an uploader dependent metadata variable doesn't change its values with the time. Using uploader dependent metadata and uploader independent web metadata, the category of web videos are classified based on the attributes such as- length of the web video, number of comments, number of views, average ratings and number of rate. We propose a novel and effective methodology to extract the metadata from web videos and classify them based on the extracted metadata by applying data mining techniques. For experimental purpose, out of the total metadata dataset, 60% are used for training and remaining 40% are used for testing the classification model built using Random Tree and J48 classification methods. The results are analyzed and the efficiency of the proposed method has been demonstrated. The system model of the proposed system is represented in Figure 1. It consists of the following components:

i) Web video metadata extraction

ii) Metadata refinement

iii) Classification model

iv) Classification analysis.

The functionality of each component of the proposed system model is discussed in the following subsections.

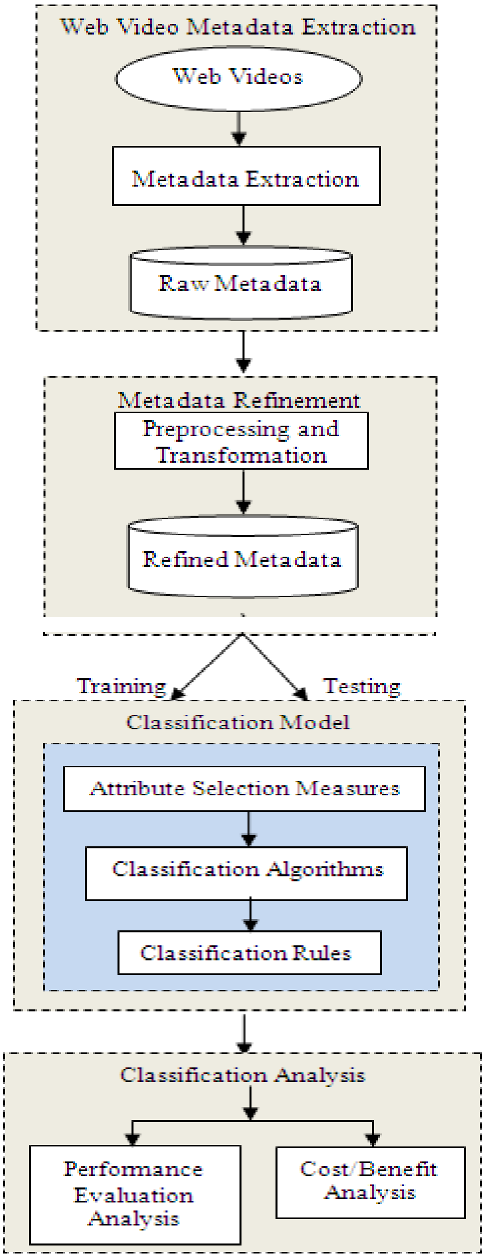


Figure 1: System model of the proposed methodology

3.1 Web Video Metadata Extraction

The web videos for a specific domain such as ‘Entertainment’, ‘News and Politics’, ‘Sports’, ‘People and Blogs’ are randomly selected and given to the InfoExtractor tool, to extract different types of web metadata. The extracted metadata will be in the form of text and these metadata are then stored in a disk [7] with ARFF or CSV file format. The Algorithm 1 is presented for the web video metadata extraction process as follows:

```
Algorithm: Web_Video_Metadata_Extraction(WV1, WV2, WV3……WVn)
Input: Web videos URL
Output: Metadata of web videos
Algorithm Steps
  1. For each selected web videos
  2. Give the selected web video URL as input to InfoExtractor
  3. Extract metadata in text format
  4. Store metadata in database
  5. End for
```

3.2 Metadata Refinement

The input to this component is raw metadata extracted from the web videos. This raw metadata has to be preprocessed for the refinement such as file format conversion and to identify the unimportant metadata. The WEKA tool is used for file preprocessing and to build classification model. The extracted raw metadata are converted to ARFF or CSV format from the text format for effective classification. Care is taken while converting to ARFF or CSV format so that all the metadata are transformed to ARFF or CSV format correctly. Some web videos might have less metadata information, whereas some web videos might have more metadata information. Through observations, it is found that, all web videos contains minimum metadata information such as- length, view counts, ratings, average ratings and number of comments, author information and URL. Among these minimum metadata information of web videos, only the numeric and nominal attributes-length, view counts, ratings, average ratings, category and number of comments are considered for classification. The refined metadata attributes are stored in a database [7] for classification.

3.3 Classification Model

This is the third component of the system model, which in turn consists of three sub-components such as,

i) Attribute selection measures

ii) Classification algorithm

iii) Classification rules.

The efficiency of the classification result largely depends on the classification model itself. Hence, construction of robust classification model plays important role in classification. The classification model construction for web videos are discussed in the following subsections.

3.3.1 Attribute Selection Measures

The attribute selection measures provide a splitting criteria for each attribute describing the given tuples. The attribute selection measures for web video metadata consisting of i) Information needed to identify the category of an element of a metadata tuple. ii) Information gain of each attributes. iii) Splitting criteria. As discussed in the section 3.2, only five numeric attributes and sixteen nominal class labels (outcomes) are considered for the dataset selected, and are listed in Table 1.

Table 1: Attribute selection for classification

S. N	***Length***	***Rate***	***Comments***	***Rating***	***Views***	***Class Labels : Category (Outcome)***
1	150	4.5	56	80	568	People and Blogs
2	140	2.5	78	56	789	Entertainment
3	450	3.6	36	20	110	Sports
⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	560	1.5	10	30	55	UNA

The procedure to measure attribute selections for the web video metadata are discussed as follows: Let D be the data set of class labeled tuples and (C_i,D) be the set of tuples of class C_i in D. Then the information needed to identify the class of an element of a data tuple is given by

$$\text{Info (D)} = -\sum_{i=1}^{M} \frac{CiD}{D} \log_2\left(\frac{CiD}{D}\right) \ldots\ldots\ldots\ (1)$$

Where **M** is the number of class labels present in the dataset. Then the formation gain of attribute ‘A’ can be measured by using the formula

$$\boxed{\text{Gain (A) = Info (D) – Info\_A(D)} \quad \ldots\ldots\ldots..(2)}$$

Where, Info_A(D) is expected information required to classify a tuple from the dataset D based on the partitioning by the attribute ‘A’ and which can be measured as

$$\boxed{\text{Info\_A(D)} = \sum_{i=1}^{n} \frac{|D_i|}{|D|} * \text{Info}(D_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots..(3)}$$

Where, D_i is the data partitions of dataset D, and Info (D_i) can be calculated by using eq (1). Using Eq(1),(2) and (3) information gain of each attribute will be calculated and the attribute which has highest information gain will be labeled as splitting node as shown in Figure 2. The splitting criterion decides which attribute to test at each node by determining the best way to split or partition the tuples into different categories. The attributes- Length of the web video, Number of comments, Number of views, Average ratings and Number of rates are continuous valued attributes. In this case, the test at node N has two possible outcomes (2 partitions) at any given time, with respect to the conditions “Attribute values<= split_point” and “Attribute values> split_point”.

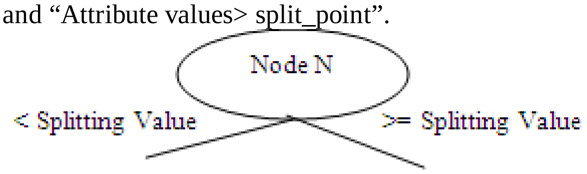


Figure 2: Labeling root node of the tree

The splitting point will be taken as mean (or mid) value of the each attribute in which values are sorted in increasing order. In the same way, the tuples are recursively partitioned and nodes are labeled with attribute which has highest information gain such that D_1 holds the subset of class-labeled tuples in D for which attribute values<=split_point, and the partition D_2 holds the rest.

3.3.2 Classification Algorithms

The proposed system model uses Decision Tree- Random Tree and J48 classification algorithms to classify/partition the web videos. Decision Tree is a tree structured predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes; the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes represents the final value (classification) of the dependent variable. The working of the Random Tree and J48 classification algorithm is described as follows:

A) Random Tree classification algorithm

Random decision tree algorithm constructs multiple decision trees randomly. While constructing each tree, the algorithm picks a "remaining" feature randomly at each node expansion. A nominal feature such as ‘Entertainment’, ‘Sports’, ‘Music’, etc is considered "remaining" if the same categorical features has not been chosen previously in a particular decision path starting from the root of the tree to the current node. Once a categorical feature is chosen, it is ineffective to pick it again on the same decision path because every sample in the same path will have the same value. However, a continuous features such as ‘Length’, ‘Comments’, ‘Ratings’ etc, can be chosen more than once in the same decision path. Each time a random threshold is selected for continuous features to split the tree.

B) J48 classification algorithm

J48 is bespoke version of C4.5 classification algorithm. The J48 algorithm generates a classification-decision tree for the web video metadata data-set by recursive partitioning the tuples. The decision tree is grown using depth-first strategy. The algorithm considers all the possible tests that can split the metadata data set and selects a test that gives the best information gain. For each continuous attributes of the web video such as ‘Length’, ‘Ratings’, ‘Comments’ etc, binary tests involving every distinct values of the attribute are considered. In order to gather the information gain of all these binary tests efficiently, the training data set belonging to the node in consideration is sorted and the information gain of the binary partition point based on each distinct values are calculated and sub trees are formed accordingly. This process is repeated for each continuous attributes.

3.3.3 Classification Rules

A part or segment of the Decision Tree structure of Random Tree (RT) and J48 classification model for the dataset chosen is represented in Figure 3.

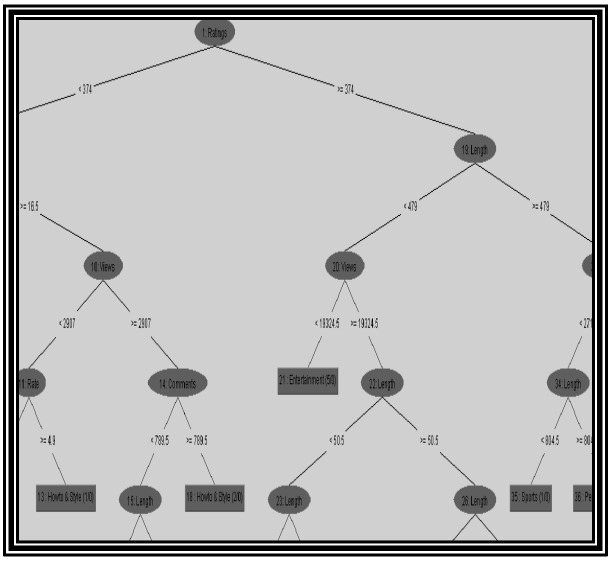
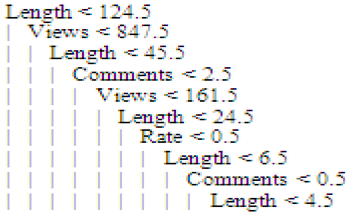
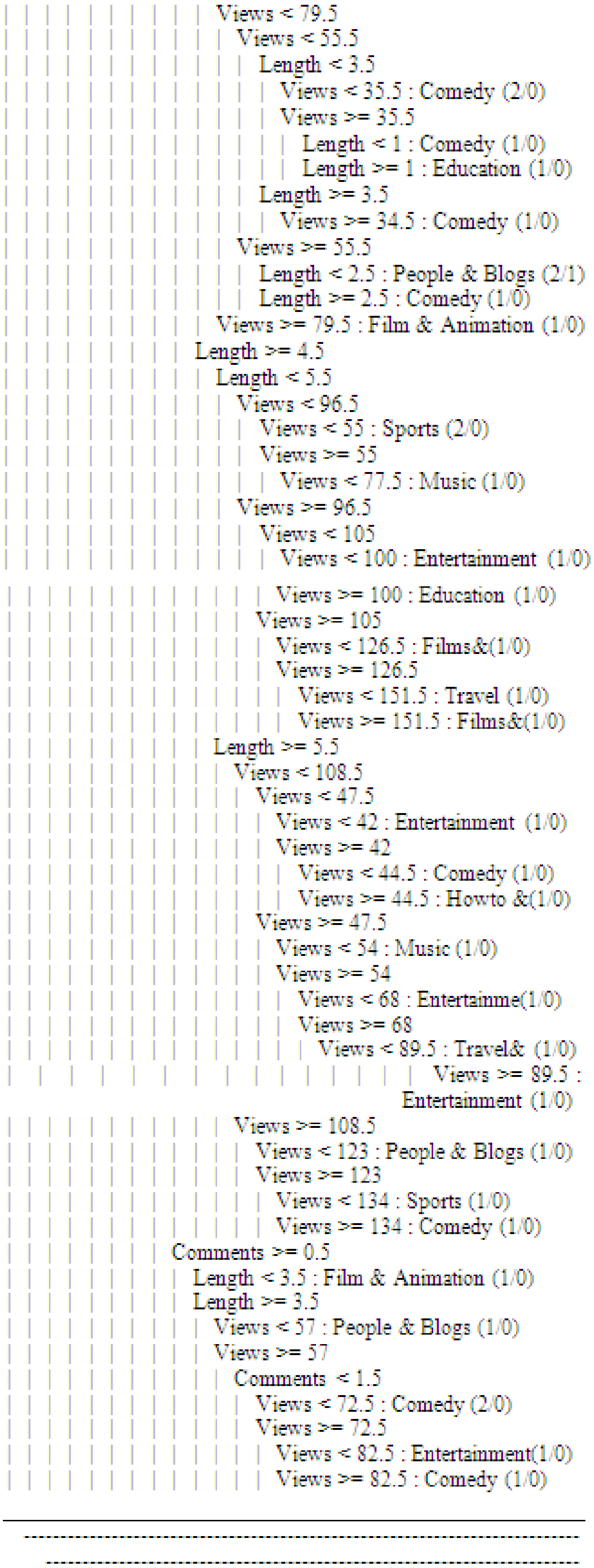


Figure 3: Tree structure result of RT and J48 classification model.

The above tree can be converted to classification rules by traversing the path from root node to each leaf node in the tree. The classification rules extracted from Figure 2 are as follows:



```
Length < 124.5
|   Views < 847.5
|   |   Length < 45.5
|   |   |   Comments < 2.5
|   |   |   |   Views < 161.5
|   |   |   |   |   Length < 24.5
|   |   |   |   |   |   Rate < 0.5
|   |   |   |   |   |   |   Length < 6.5
|   |   |   |   |   |   |   |   Comments < 0.5
|   |   |   |   |   |   |   |   |   Length < 4.5
```



```
| | | | | | | | | | Views < 79.5
| | | | | | | | | | | Views < 55.5
| | | | | | | | | | | | Length < 3.5
| | | | | | | | | | | | | Views < 35.5 : Comedy (2/0)
| | | | | | | | | | | | | Views >= 35.5
| | | | | | | | | | | | | | Length < 1 : Comedy (1/0)
| | | | | | | | | | | | | | Length >= 1 : Education (1/0)
| | | | | | | | | | | | Length >= 3.5
| | | | | | | | | | | | | Views >= 34.5 : Comedy (1/0)
| | | | | | | | | | | Views >= 55.5
| | | | | | | | | | | | Length < 2.5 : People & Blogs (2/1)
| | | | | | | | | | | | Length >= 2.5 : Comedy (1/0)
| | | | | | | | | | Views >= 79.5 : Film & Animation (1/0)
| | | | | | | | | Length >= 4.5
| | | | | | | | | | Length < 5.5
| | | | | | | | | | | Views < 96.5
| | | | | | | | | | | | Views < 55 : Sports (2/0)
| | | | | | | | | | | | Views >= 55
| | | | | | | | | | | | | Views < 77.5 : Music (1/0)
| | | | | | | | | | | Views >= 96.5
| | | | | | | | | | | | Views < 105
| | | | | | | | | | | | | Views < 100 : Entertainment  (1/0)

| | | | | | | | | | | | | Views >= 100 : Education  (1/0)
| | | | | | | | | | | | Views >= 105
| | | | | | | | | | | | | Views < 126.5 : Films&(1/0)
| | | | | | | | | | | | | Views >= 126.5
| | | | | | | | | | | | | | Views < 151.5 : Travel  (1/0)
| | | | | | | | | | | | | | Views >= 151.5 : Films&(1/0)
| | | | | | | | | | Length >= 5.5
| | | | | | | | | | | Views < 108.5
| | | | | | | | | | | | Views < 47.5
| | | | | | | | | | | | | Views < 42 : Entertainment  (1/0)
| | | | | | | | | | | | | Views >= 42
| | | | | | | | | | | | | | Views < 44.5 : Comedy (1/0)
| | | | | | | | | | | | | | Views >= 44.5 : Howto &(1/0)
| | | | | | | | | | | | Views >= 47.5
| | | | | | | | | | | | | Views < 54 : Music (1/0)
| | | | | | | | | | | | | Views >= 54
| | | | | | | | | | | | | | Views < 68 : Entertainme(1/0)
| | | | | | | | | | | | | | Views >= 68
| | | | | | | | | | | | | | | Views < 89.5 : Travel&  (1/0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Views >= 89.5 :
                                         Entertainment  (1/0)
| | | | | | | | | | | Views >= 108.5
| | | | | | | | | | | | Views < 123 : People & Blogs (1/0)
| | | | | | | | | | | | Views >= 123
| | | | | | | | | | | | | Views < 134 : Sports (1/0)
| | | | | | | | | | | | | Views >= 134 : Comedy (1/0)
| | | | | | | | Comments >= 0.5
| | | | | | | | | Length < 3.5 : Film & Animation (1/0)
| | | | | | | | | Length >= 3.5
| | | | | | | | | | Views < 57 : People & Blogs (1/0)
| | | | | | | | | | Views >= 57
| | | | | | | | | | | Comments  < 1.5
| | | | | | | | | | | | Views < 72.5 : Comedy (2/0)
| | | | | | | | | | | | Views >= 72.5
| | | | | | | | | | | | | Views < 82.5 : Entertainment(1/0)
| | | | | | | | | | | | | Views >= 82.5 : Comedy (1/0)
```

3.4 Classification Analysis

In this section, performance evaluation measures such as TP, FP, precision, recall and F-Measure will be calculated to measure classification accuracy and efficiency of RT and J48 classification model. Also the classification accuracy of RT and J48 will be compared. To analyze the cost/benefit of each web video category, the result of classification model with highest accuracy will be taken. In the cost/benefit analysis, the minimum cost and maximum cost of each web video category with classification accuracy will be represented.

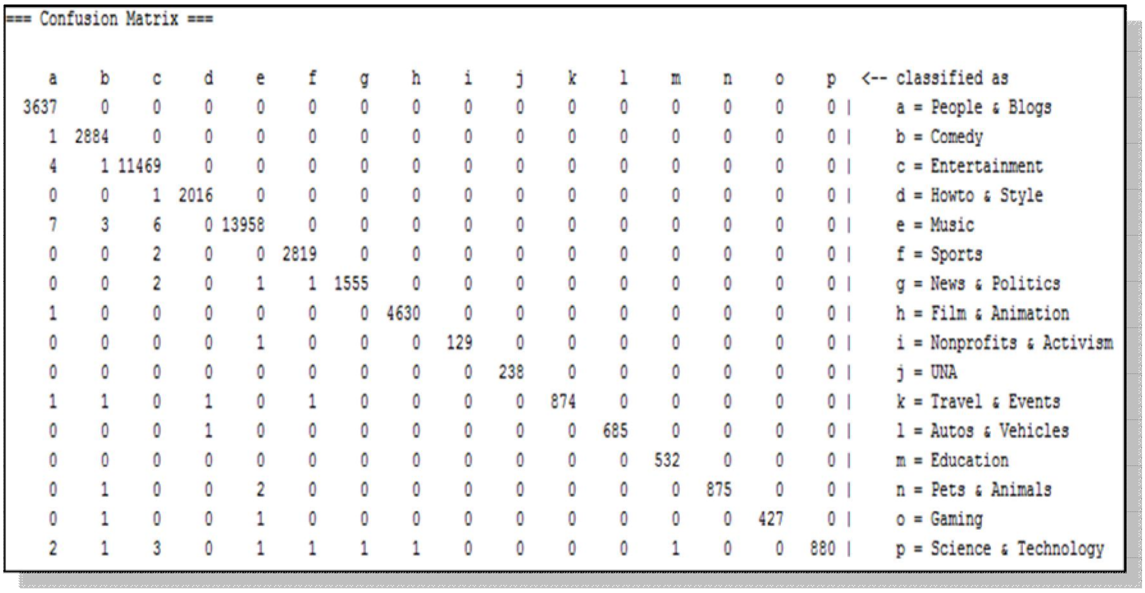
4. Experimental Results and Discussions

4.1 Classification using Random Tree and J48 classification model

To test the efficiency of the classification models constructed using Random tree and J48, the dataset is downloaded from the website [8] which consists of 47660 web video metadata instances. The performance of the model is measured in terms of number of correctly classified instances, number of incorrectly classified instances, TP rate, FP rate, precision, recall and F-score. The Table 2 represents classification result obtained by the Random Tree classification model.

Table 2: Classification result of Random Tree classification model

Sl.No	***Class Labels***	***Total Weight***	***Correctly Classified***	***Incorrectly Classified***	***TP***	***FP***	***Precision***	***Recall***	***F-Measure***
1	People&Blog	3637	3637	0	1	0	0.996	1	0.998
2	Comedy	2885	2884	1	1	0	0.997	1	0.998
3	Entertainment	11474	11469	5	1	0	0.999	1	0.999
4	Howto&Style	2017	2016	1	1	0	0.999	1	0.999
5	Music	13974	13958	16	0.999	0	1	0.999	0.999
6	Sports	2821	2819	2	0.999	0	0.995	0.999	0.999
7	News	1559	1555	4	0.997	0	0.999	0.997	0.998
8	Films&Animation	4631	4630	1	1	0	1	1	1
9	Non-Profit	130	129	1	0.992	0	1	0.992	0.996
10	UNA	238	238	0	1	0	1	1	1
11	Travel&Events	878	874	4	0.995	0	1	0.995	0.998
12	Autos&Vehicles	686	685	1	0.999	0	1	0.999	0.998
13	Education	532	532	0	1	0	0.998	1	0.998
14	Pets&Animals	878	875	3	0.997	0	1	0.997	0.999
15	Gaming	429	427	2	0.988	0	1	0.995	0.997
16	Science&Tech	891	880	11	0.999	0	1	0.988	0.994
	Total	***47660***	***47608***	***52***	***0.999***	***0.000***	***0.995***	***0.995***	***0.995***



```
=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i     j     k     l     m     n     o     p   <-- classified as
  3637     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     a = People & Blogs
     1  2884     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     b = Comedy
     4     1 11469     0     0     0     0     0     0     0     0     0     0     0     0     0 |     c = Entertainment
     0     0     1  2016     0     0     0     0     0     0     0     0     0     0     0     0 |     d = Howto & Style
     7     3     6     0 13958     0     0     0     0     0     0     0     0     0     0     0 |     e = Music
     0     0     2     0     0  2819     0     0     0     0     0     0     0     0     0     0 |     f = Sports
     0     0     2     0     1     1  1555     0     0     0     0     0     0     0     0     0 |     g = News & Politics
     1     0     0     0     0     0     0  4630     0     0     0     0     0     0     0     0 |     h = Film & Animation
     0     0     0     0     1     0     0     0   129     0     0     0     0     0     0     0 |     i = Nonprofits & Activism
     0     0     0     0     0     0     0     0     0   238     0     0     0     0     0     0 |     j = UNA
     1     1     0     1     0     1     0     0     0     0   874     0     0     0     0     0 |     k = Travel & Events
     0     0     0     1     0     0     0     0     0     0     0   685     0     0     0     0 |     l = Autos & Vehicles
     0     0     0     0     0     0     0     0     0     0     0     0   532     0     0     0 |     m = Education
     0     1     0     0     2     0     0     0     0     0     0     0     0   875     0     0 |     n = Pets & Animals
     0     1     0     0     1     0     0     0     0     0     0     0     0     0   427     0 |     o = Gaming
     2     1     3     0     1     1     1     1     0     0     0     0     1     0     0   880 |     p = Science & Technology
```

Figure 4: Confusion matrix of RT classification result

It is observed from the experimental result that, out of 47660 instances, 47608 tuples are correctly classified and 52 are incorrectly classified by the Random tree classification model. The class labels 'Music', 'Films and Animation', 'Non-Profit', 'UNA', 'Travels and Events', 'Autos and Vehicles', 'Pets and Animals', 'Gaming', and 'Science and Technology' has highest precision among class labels used. The overall efficiency of the Random tree classification model is 99.5%. The classification results of RT classification model are represented in the form of confusion matrix in Figure 4.

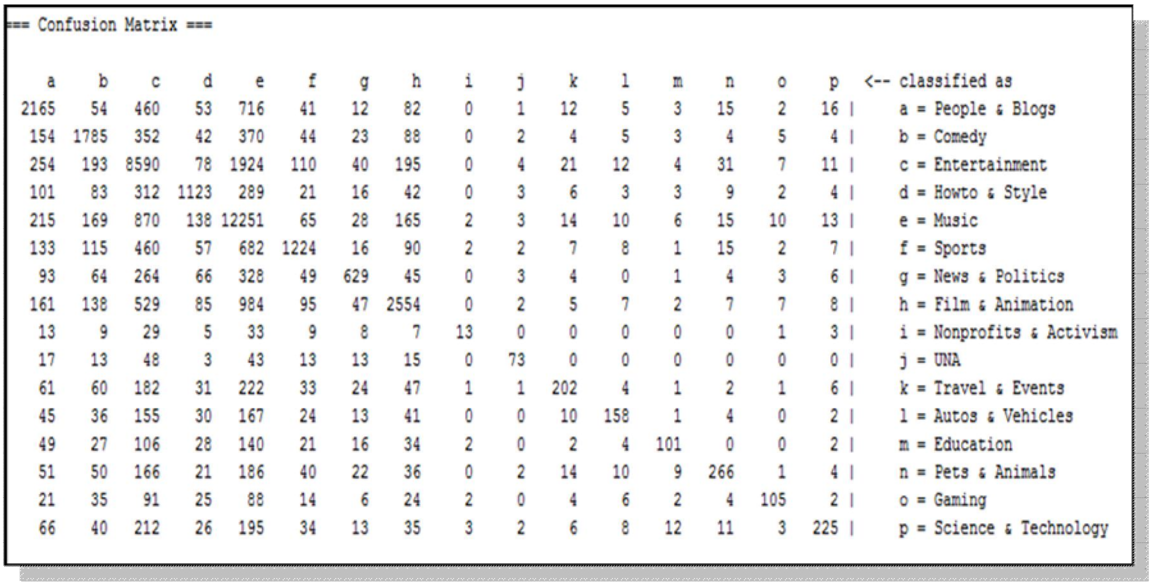
In this confusion matrix, the column 'a' and the row 'a' corresponds to the class label 'People and Blogs', similarly, the column 'b' and row 'b' corresponds to the class label 'Comedy' and so on. From the result of confusion matrix, one metadata tuple with class label 'Comedy' is misclassified as 'People and Blogs', and one tuple of 'Entertainment category is misclassified as 'Comedy' and four tuples are misclassified as 'People and Blogs'. According to the same aspect, the 'Music' category has highest misclassified tuples among remaining web video categories. The class labels 'People and Blog', 'UNA' and 'Education' are classified more effectively as compared to remaining class labels. Also the misclassification rate of the RT classification model is very less.

Similarly, the same dataset is applied to J48 classification algorithm and obtained the following result as described in the Table 3 which contains classification performance evaluation metrics such as- number of correctly classified instances, number of incorrectly classified instances, TP rate, FP rate, precision, recall and F-score.

Table 3: Classification result of J48 classification model

Sl.No	*Class Labels*	*Total Weight*	*Correctly Classified*	*Incorrectly Classified*	*TP*	*FP*	*Precision*	*Recall*	*F-Measure*
1	People&Blog	3637	2165	1472	0.595	0.033	0.602	0.595	0.598
2	Comedy	2885	1785	1100	0.619	0.024	0.622	0.619	0.62
3	Entertainment	11474	8590	2884	0.749	0.117	0.67	0.749	0.707
4	Howto&Style	2017	1123	894	0.557	0.015	0.62	0.557	0.587
5	Music	13974	12251	1723	0.877	0.189	0.658	0.877	0.752
6	Sports	2821	1224	1597	0.434	0.014	0.666	0.434	0.526
7	News	1559	629	930	0.403	0.006	0.679	0.403	0.506
8	Films&Animation	4631	2554	2077	0.552	0.022	0.73	0.552	0.628
9	Non-Profit	130	13	117	0.1	0	0.52	0.1	0.168
10	UNA	238	73	165	0.307	0.001	0.745	0.307	0.435
11	Travel&Events	878	202	676	0.23	0.002	0.65	0.23	0.34
12	Autos&Vehicles	686	158	528	0.23	0.002	0.658	0.23	0.341

13	Education	532	101	431	0.19	0.001	0.678	0.19	0.297
14	Pets&Animals	878	266	612	0.303	0.003	0.687	0.303	0.451
15	Gaming	429	105	324	0.245	0.001	0.705	0.245	0.363
16	Science&Tech	891	225	666	0.253	0.002	0.719	0.253	0.374
	Total	***47660***	***31624***	***16396***	***0.66***	***0.092***	***0.663***	***0.66***	***0.642***



```
=== Confusion Matrix ===

    a     b     c     d     e     f     g     h     i     j     k     l     m     n     o     p   <-- classified as
 2165    54   460    53   716    41    12    82     0     1    12     5     3    15     2    16 |     a = People & Blogs
  154  1785   352    42   370    44    23    88     0     2     4     5     3     4     5     4 |     b = Comedy
  254   193  8590    78  1924   110    40   195     0     4    21    12     4    31     7    11 |     c = Entertainment
  101    83   312  1123   289    21    16    42     0     3     6     3     3     9     2     4 |     d = Howto & Style
  215   169   870   138 12251    65    28   165     2     3    14    10     6    15    10    13 |     e = Music
  133   115   460    57   682  1224    16    90     2     2     7     8     1    15     2     7 |     f = Sports
   93    64   264    66   328    49   629    45     0     3     4     0     1     4     3     6 |     g = News & Politics
  161   138   529    85   984    95    47  2554     0     2     5     7     2     7     7     8 |     h = Film & Animation
   13     9    29     5    33     9     8     7    13     0     0     0     0     0     1     3 |     i = Nonprofits & Activism
   17    13    48     3    43    13    13    15     0    73     0     0     0     0     0     0 |     j = UNA
   61    60   182    31   222    33    24    47     1     1   202     4     1     2     1     6 |     k = Travel & Events
   45    36   155    30   167    24    13    41     0     0    10   158     1     4     0     2 |     l = Autos & Vehicles
   49    27   106    28   140    21    16    34     2     0     2     4   101     0     0     2 |     m = Education
   51    50   166    21   186    40    22    36     0     2    14    10     9   266     1     4 |     n = Pets & Animals
   21    35    91    25    88    14     6    24     2     0     4     6     2     4   105     2 |     o = Gaming
   66    40   212    26   195    34    13    35     3     2     6     8    12    11     3   225 |     p = Science & Technology
```

Figure 5: Confusion matrix of result of J48 classification model

The number of correctly classified instances are 31464 and 16196 instances are misclassified. The performance evaluation metrics of the class labels ‘Non Profit’, ‘

results of J48 classification model and confusion matrix is represented in Figure 5.

The class labels ‘Music’, ‘Comedy, ‘Entertainment’ and ‘Films and Animation’ are classified more effectively as compared to remaining class labels. Also the misclassification rate of the J48 classification model is more as compared to RT classification model result. From the result of confusion matrix, the majority of the misclassification rate are distributed among the class labels such as ‘People and Blogs’, ‘Comedy’, ‘Entertainment’, ‘How to and Style’, ‘Music’, ‘Sports’, ‘News’, and ‘Films and Animation’.

In this experiment, the web video metadata sets partitioned and classified effectively on random tree classification model. The comparison result is represented in Table 4.

Table 4: Comparison of classification Accuracy

Classification model	***Correctly Classifie d***	***Incorrectly classified***	***Accuracy***
Random Tree	47432	228	99.5%
J48 Classifier	31464	16196	66.02%

The experimental result shows that, RT classification model works well on numeric/continuous data as

Travel and Events’, ‘Autos and Vehicles’, and ‘Education’ are very less as compared to remaining class labels.The

compared to J48 classification model. The web video metadata datasets contains all independent attribute as continuous values. Due to this factor the J48 classification has less accuracy than RT classification model. The partition analysis of independent attributes is analyzed using Random Tree classification model in the subsection 4.2.

4.2 Attribute analysis for web video class labels

This section describes comments, rate, ratings, length and view count analysis for all identified 16 web video class labels.

The Figure 6 represents statistical comparison of attribute analysis of all the web video class labels where, X-axis represents class labels of web videos, and Y-axis represents normalized average values of - comments, rate, ratings, length and view counts of each web video class labels. To plot the graph, the values of each attributes are normalized using ‘Min-Max normalization’ method and scaled between 0 to 100. From the comparison statistics, the class labels ‘Music’, ‘Films and Animation’, ‘Entertainment’, ‘Comedy’ and ‘Sports’ as the top 5 web video class labels which are having highest number of comments. The rate information analysis of web videos has two parts- the analysis of web videos using number of ratings of the web videos, and the ratings (0 to 5) given by the online users for web videos. Also the highest number of average rate and ratings obtained by the class labels are - 'Music', 'Films and Animation', 'Comedy' and 'Entertainment' respectively.

In the length analysis of large scale web videos, we found the following top 5 web video class labels which are having large number of length units – 'Entertainment', 'Science and Technology', 'News and Politics', 'People and Blogs', 'Music' and 'Films and Animation' respectively.

In the analysis of 'views' counts of large scale web videos, statistical measurements are made to find average view count of each web video class labels, and obtained the class labels 'Music', 'Entertainment', 'Films and Animation', 'Comedy', and 'Sports' as top 5 popular web video categories respectively.

The attribute analysis of web videos shows that, nowadays, the web videos of - Entertainment, Movie, Music and Sports class labels are highly attracting the online users as compared to other class labels web videos.

4.3 Cost/Benefit analysis of web video categories using RT classification model

By the comparison of results of J48 classification model and RT classification model, the RT classification model is found with highest accuracy. Hence, this section represents cost/benefit analysis of only RT classification model for each 16 nominal category of web videos. The minimization of cost/benefit and its percentage of classification accuracy are represented in Table 5.

The experimental result shows that, the RT classification method is the most effective way to classify the web videos based on their category. Also the result of classification, attribute analysis and cost benefit analysis of web videos shows that, there are only few category of web videos are most popularly being used by the online users and are rapidly sharing on social websites. By the results of classification/partitioning and cost benefit analysis, we found the following top 5 web video categories - 'Films and Animation', 'Music', 'Entertainment', 'Comedy', 'Sports' and 'News and Politics' respectively.

Table 5: Cost/Benefit analysis of each category using Random Tree classification model.

Class/Category	Minimum Cost/Benefit	% of Classification Accuracy
People&Blog	53	99.88
Comedy	9	99.98
Entertainment	19	99.96
Howto&Style	3	99.99
Music	22	99.95
Sports	5	99.98
News	5	99.98
Films&Animation	2	99.99
Non-Profit	1	99.99
UNA	4	99.99
Travel&Events	1	99.99
Autos&Vehicles	1	99.99
Education	1	99.99
Pets&Animals	3	99.99
Gaming	2	99.99
Science&Tech	3	99.97

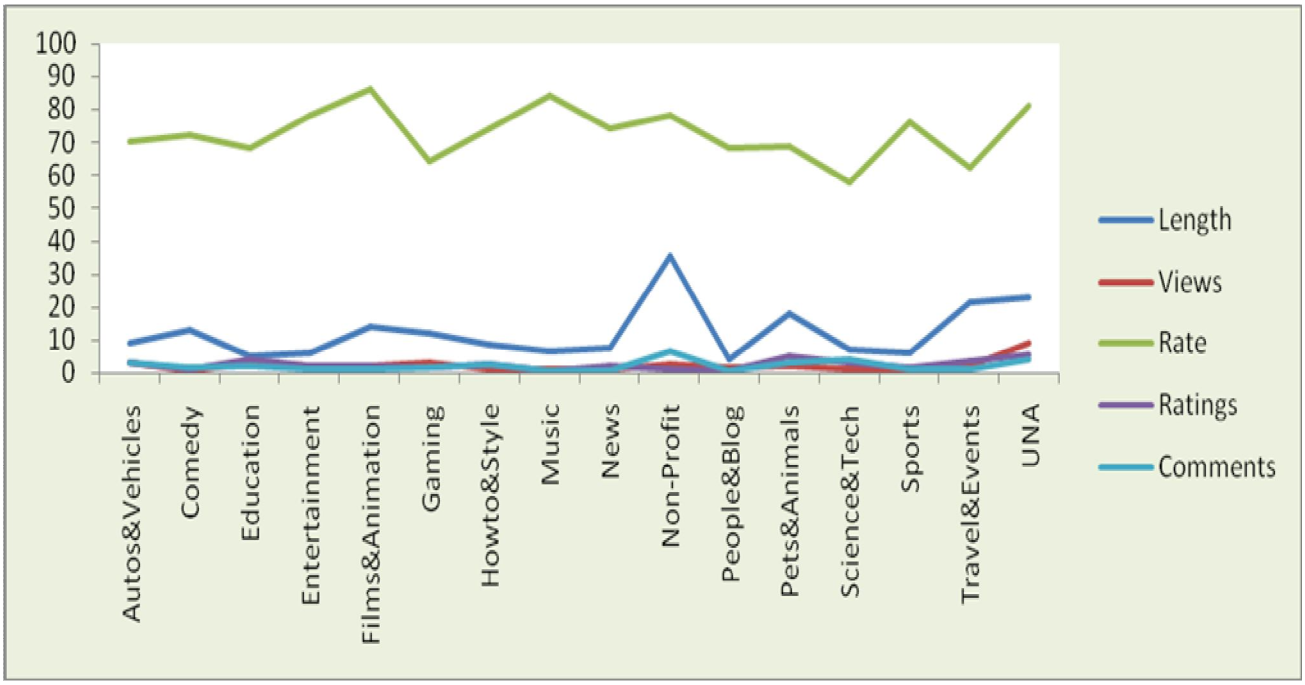


Figure 6: Attribute analysis of large scale web videos

5.Conclusion and Future Works

In this work, we classified/partitioned web videos based on their category using web metadata. The web video metadata are extracted and stored in a database for classification. The Random Tree (RT) and J48 classification algorithms are chosen to classify/partitioning the web videos. The classification results of RT and J48 classification models are compared and found RT classification model is more efficient for classify web videos using metadata. Also, category wise cost/benefit analysis of RT classification result is analyzed and minimum cost/benefit and maximum cost/benefit are found with classification accuracy. Difficulties were arrived during the classification due to insufficient web metadata. Only 79% tuples were classified and 21% tuples are ignored by the Random Tree and J48 classification models due to poor tuple records. Also the J48 classification model has less efficiency on partitioning web video categories based in independent attributes. This is because, the independent attribute values are numeric. The future work is to improve the classification accuracy of J48 classification model on the classification/partition of web videos.

References

[1] Yang Song, Ming Zhao, Jay Yagnik, and Xiaoyun Wu, “Taxonomic Classification for Web-based Videos”, Google Inc., Mountain View, CA 94043, USA

[2] Siddu P. Algur, Prashant Bhat, Suraj Jain, “Metadata Construction Model for Web Videos: A Domain Specific Approach”, International Journal of Engineering and Computer Science, December 2014.

[3] Roger J. Lewis, “Introduction to CART”, Annual Meeting of the Society for Academic Emergency Medicine, 2000.

[4] John R. Zhang, Yang Song, and Thomas Leung, “Improving Video Classification via YouTube Video Co-Watch Data”, SBNMA’11, December 1, 2011.

[5] Chunneng Huang, Tianjun Fu and Hsinchun Chen, “Text-based Video Content Classification for Online Video-sharing Sites”, http://crcv.ucf.edu/projects/gist3d

[6] Zheshen Wang, Ming Zhao, Yang Song, Sanjiv Kumar, and Baoxin Li, “YouTubeCat: Learning to CategorizeWildWeb Videos”, Google Research.

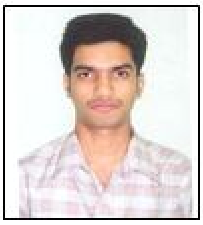
[7] Siddu P. Algur, Prashant Bhat, Suraj Jain, “The Role of Metadata in Web Video Mining: Issues and Perspectives”, International Journal of Engineering Sciences & Research Technology, February-2015.

[8] Statistics and Social Network of YouTube Videos : http://netsg.cs.sfu.ca/youtubedata/

AUTHOR

Dr. Siddu P. Algur is working as Professor, Dept. of Computer Science, Rani Channamma University, Belagavi, Karnataka, India. He received B.E. degree in Electrical and Electronics from Mysore University, Karnataka, India, in 1986. He received his M.E. degree in from NIT, Allahabad, India, in 1991.He obtained Ph.D. degree from the Department of P.G. Studies and Research in Computer Science at Gulbarga University, Gulbarga.

He worked as Lecturer at KLE Society’s College of Engineering and Technology and worked as Assistant Professor in the Department of Computer Science and Engineering at SDM College of Engineering and Technology, Dharwad. He was Professor, Dept. of Information Science and Engineering, BVBCET, Hubli, before holding the present position. He was also Director, School of Mathematics and Computing Sciences, RCU, Belagavi. He was also Director, PG Programmes, RCU, Belagavi. His research interest includes Data Mining, Web Mining, Big Data and Information Retrieval from the web and Knowledge discovery techniques. He published more than 45 research papers in peer reviewed International Journals and chaired the sessions in many International conferences.



Mr. Prashant Bhat is pursuing Ph.D programme in Computer Science at Rani Channamma University Belagavi, Karnataka, India. He received B.Sc and M.Sc (Computer Science) degrees from Karnatak University, Dharwad, Karnataka, India, in 2010 and 2012 respectively. His research interest includes Data Mining, Web Mining, web multimedia mining and Information Retrieval from the web and Knowledge discovery techniques, and published 7 research papers in International Journals. Also he has attended and participated in International and National Conferences and Workshops in his research field.