Inner Classification of Clusters for Online News

Harmandeep Kaur[1,] Sheenam Malhotra[2]
[1](Computer Science and Engineering Department, Shri Guru Granth Sahib World University Fatehgarh Sahib)
[2](Assistant Professor, Computer Science and Engineering Department, Shri Guru Granth Sahib World University Fatehgarh Sahib)
harmandip201@gmail.com

ABSTRACT

Data mining is a term which proceeds with large amount of data. Online news classification is vast area of data mining. As large numbers of articles are published it is time consuming task to select the interesting one. In the previous research work the manual system were into action as the people were used to extract the news manually. A lot of research work has been already done in this field. So in research work the method of news classification is proposed define a hybrid model for the inner classification of clusters. It divide every cluster into sub clusters with the help of K mean, CART, SVM, HMM algorithms.

Keywords*:*

Hidden Markov Model(HMM),Support Vector Machine(SVM),K Mean, CART

I. INTRODUCTION

Data mining is process of discovering interesting knowledge such as patterns, associations, changes, anomalies and significant structures, from large amounts of data stored in database, data warehouse, or other information repositories. Due to the wide availability of huge amount of data in electronic form, and imminent need for turning such data into useful information and knowledge for broad application including market analysis, business management and decision support, data mining has attracted a great deal of attention in information industry in recent year. Data mining has been popularly treated as synonym of knowledge discovery in database, although some researchers view data mining as an essential step of knowledge discovery. Since the emergence of WWW , it is essential to handle a very large amount of electronic data of which majority is in the form of text. This can be handle by various data mining technique[2].

A knowledge discovery process consist of an iterative sequence of following steps[1]:

- *Data cleaning,* which handle noisy, erroneous, missing, or irrelevant data.
- *Data integration*, where multiple, heterogeneous data source may be integrated into one.
- *Data selection,* where data relevant to analysis task are retrieved from database.
- *Data transformation,* where data are transformed or consolidated into form appropriate for mining by performing aggregate operations.
- *Data mining,* which is essential process where intelligent methods are applied in order to extract data patterns.

- *Pattern evaluation,* which is to identify the truly interesting patterns representing knowledge based on some interestingness measure.
- *Knowledge presentation,* where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

II. TEXT CLASSIFICATION

Mostly the information is stored in the form of text like e mails, web pages, newspaper article, market research reports, complaint letter from customer and internally generated reports. Online news papers provide news under various categories like national, international, politics, finance, sports, entertainment etc[2]. News article on topical issue are helpful for company manager and other decision maker. It is time consuming task to select the interesting one from large amount of news article . With the help of news categorization we obtain the information quickly. Text classification is also an important part of text mining[3] . Text classification is based on expert knowledge and how to classify the document under the given set of categories. Data mining classification start with training set of documents that are already labeled with class. Text classification has two flavours as single label and multi label[3]. A single label document belong to only one class and multi label document may belong to more than one class.

III. CLUSTERING

Document clustering has become an increasingly important technique for unsupervised document organization, automatic topic extraction, and fast information retrieval or filtering. For example, a web search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories, as is achieved by search engines such as Northern Light and Vivisimo. Similarly, a large database of documents can be pre-clustered to facilitate query processing by searching only the cluster that is closest to the query[10].

IV. PROPOSED WORK

Figure no 1 represent the process of news classification .

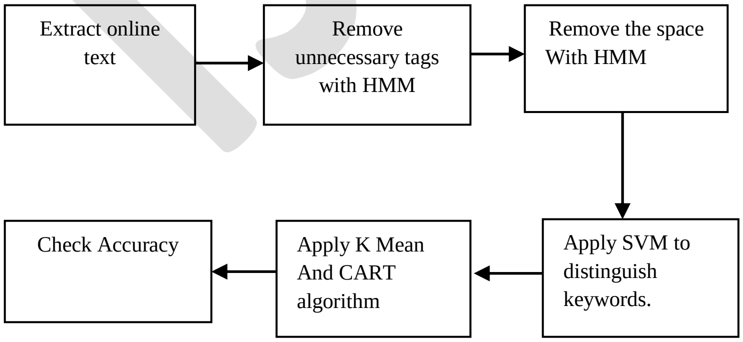

Extract online text
Remove unnecessary tags with HMM
Remove the space With HMM
Apply SVM to distinguish keywords.
Apply K Mean And CART algorithm
Check Accuracy

Figure no. 1 : News Classification Process

HMM: HMM is Hidden Markov Model, In this work HMM is used for text extraction. When given the URL address of any newspaper, Then source code will be displayed it is not in proper text form. Some html tags are also include in this source code. So first of all remove all these tags from source code . When html tags are removed then empty space is shown in the place of tags. Then next step of HMM is to remove this empty space . With this we obtain text in proper form. This text is used further as an input in SVM for classification.

SVM: SVM is Support Vector Machine , It is a binary classifier used for text classification . It is based on structure risk minimization principal. Positive data is represent as 1 and negative data represent as 0. It is used to distinguished the keywords.

K MEAN: K mean is used to create the clusters . It grouping the data into K clusters. The main aim of K mean is to minimize the Euclidean distance between data points . It create the clusters of different categories like clustering for sports is performed as hockey, football, cricket etc . For matrimonial grouping the data related with hindu matrimonial and muslim matrimonial and Clustering for entertainment is bollywood and hollywood.

CART: CART is classification and regression tree. It is used to set the counter with higher value . it create a hierarchy for the classification of each news.

V. EXPERIMENTAL RESULT

To evaluate the effectiveness of news classification method proposed in this paper , we collect the news from different different newspapers related wit sports , entertainment and matrimonial as shown in table 1.

CATEGORIES	SUBCATEGORIES	NEWS PROCESSED
ENTERTAINMENT	BOLLYWOOD HOLLYWOOD	31 26
MATRIMONIAL	HINDU MUSLIM	35 26
SPORTS	CRICKET FOOTBALL HOCKEY	36 34 38

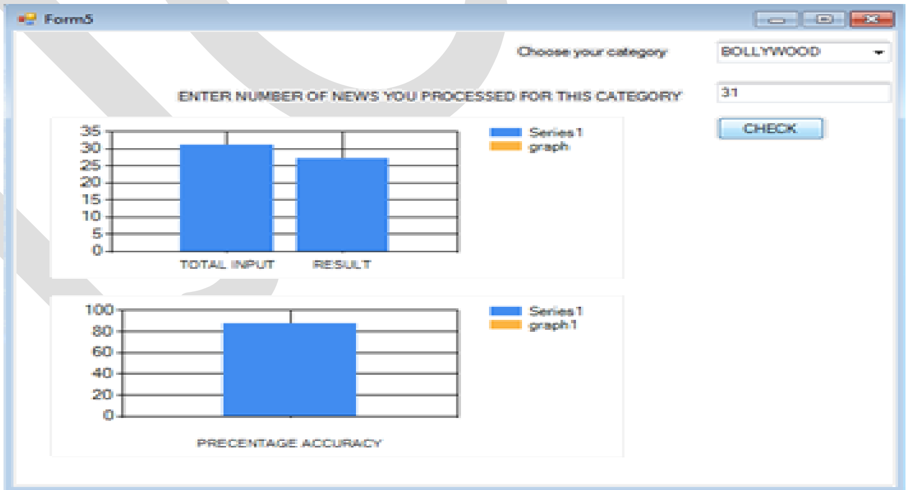
Table 1. Processed News Of Different Categories.

Total input news related with bollywood is 31 but it classify 27 news correctly. Total Hollywood news as input is 26 and it classify 22 news correctly. Among the 35 hindu matrimonial news it classify 31 news correctly. Input news of muslim matrimonial is 26 but 20 news classify correctly. In sports total cricket news as input is 36 and 32 classify correctly. Out of 34 football news 29 news classify correctly. And for hockey out of 38 input news 35 news classify correctly. Table 2 shows the percentage of accuracy for different news.

CATEGORIES	SUBCATEGORIES	ACCURACY
ENTERTAINMENT	**BOLLYWOOD** **HOLLYWOOD**	**87.09%** **84.61%**
MATRIMONIAL	**HINDU** **MUSLIM**	**88.57%** **76.92%**
SPORTS	**CRICKET** **FOOTBALL** **HOCKEY**	**88.88%** **85.29%** **92.10%**

Table 2. Percentage Of Accuracy Of Different News.

Figure 2 represent the total processed bollywood news and percentage of accuracy. Total processed bollywood news are 31 and 27 news are classify correctly and accuracy is 87.09%.

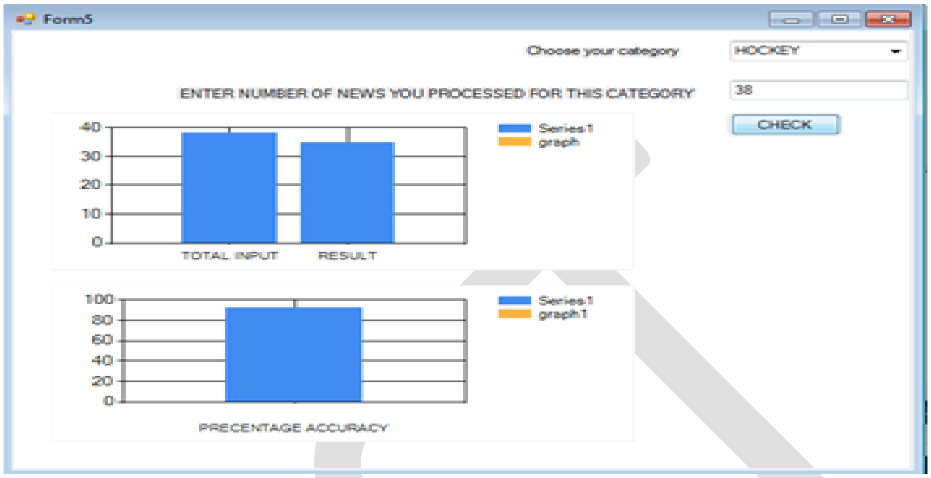


Choose your category BOLLYWOOD

ENTER NUMBER OF NEWS YOU PROCESSED FOR THIS CATEGORY 31

Figure 2. Accuracy Of bollywood News.

Figure 3 represent the processed hockey news and percentage of accuracy. Total input is 38 and 35 news classify correctly then the accuracy is 92.10%.

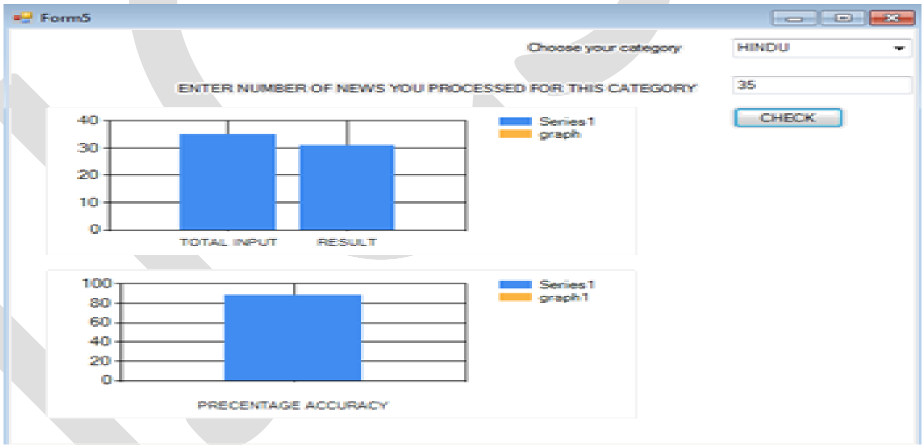


Choose your category: HOCKEY

ENTER NUMBER OF NEWS YOU PROCESSED FOR THIS CATEGORY: 38

Figure 3. Accuracy Of hockey News

Figure 4 represent the processed hindu matrimonial news . Total input is 35 news and it classify 31 news correctly then the percentage of accuracy is 88.57%



Choose your category: HINDU

ENTER NUMBER OF NEWS YOU PROCESSED FOR THIS CATEGORY: 35

Figure 4. Accuracy of hindu matrimonial news

Same as figures 2,3,4 we check the accuracy of other categories like hollywood, football, cricket and muslim .The percentage of accuracy of all these categories are given in table 2.

VI.CONCLUSION AND FUTURE SCOPE

A new hybrid approach is developed and experimented with online news for classification. Different categories like Sports, Entertainment, Matrimonial are classified into subcategories successfully . HMM and SVM used with K mean and CART algorithms provide good result for inner cluster classification of online news . It provide help to take any decision quickly with because mostly textual information is available online. Effective retrieval is difficult without good classification of documents. Experimental result show the performance of this approach in the form of accuracy . In the future a neural network is used to optimize the code.

References

[1] Mr.S.P Deshpande and Dr. V.M Thakre," data mining system and applications:a review,"international journal of distributed and parallel system(IJDPS) Vol.1,No.1, September 2010.

[2] Krishnalal G, S Babu Rengarajan, K G Srinivasagan ,"A new text mining approach based on HMM -SVM for web news classification" International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 19,2010.

[3] Vandana Korde, C Namrata Mahender,"Text classification and classifier a survey," International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[4] Daniel I. Morariu, Lucian N. Vintan, and Volker Tresp,"Meta-Classification using SVM Classifiers for Text Documents,"World Academy of Science, Engineering and Technology 21 2006.

[5] D. Morariu, R. Cre¸tulescu and L. Vin¸tan,"Improving a SVM Meta-classifier for Text Documents by using Naïve-Bayes,"Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844.

[6] Mita K. Dalal, Mukesh A.Zaveri,"Automatic text classification," International Journal of Computer Applications (0975 – 8887) Volume 28– No.2, August 2011.

[7] Rama Bharath Kumar, Bangari Shravan Kumar, Chandragiri Shiva Sai Prasad,"financial news classification using SVM," International Journal of Scientific and Research Publications, Volume 2, Issue 3, March 2012 .

[8] Chee-Hong Chan Aixin Sun Ee-Peng Lim," Automated Online News Classification with Personalization,"4th international conference on asian digital libraries , 12-2001.

[9] Thorsten Joachims,"Text Categorization With Support Vector Machines: Learning with many relevant features,"technical report 23, universitat Dortmund, LS VIII, 1997.

[10] shri zhong and joydeep ghosh ," A Comparative study of generative models for document clustering," the university o texas at Austin ,TX 78712-1084,Vol 4, No1,2008.

[11] yiming yang and xin liu,"A re-examination of text categorization methods," caenegie mellon university pittsburg , PA 15213-3702 , USA ,Vol 18, No.2, March 2012.