International Journal of Engineering and Techniques - Volume 2 Issue 3, May – June 2016

RESEARCH ARTICLE OPEN ACCESS

Automatic Text Categorization on News Articles

Muthe Sandhya[1], Shitole Sarika[2], Sinha Anukriti[3], Aghav Sushila[4]

1,2,3,4(Department of Computer, MITCOE, Pune, India)

Abstract:

Text categorization is a term that has intrigued researchers for quite some time now. It is the concept in which news articles are categorized into specific groups to cut down efforts put in manually categorizing news articles into particular groups. A growing number of statistical classification and machine learning technique have been applied to text categorization. This paper is based on the automatic text categorization of news articles based on clustering using k-mean algorithm. The goal of this paper is to automatically categorize news articles into groups. Our paper mostly concentrates on K-mean for clustering and for term frequency we are going to use TF-IDF dictionary is applied for categorization. This is done using mahaout as platform.

Keywords — **Text categorization, Preprocessing, K-mean, TF-IDF etc.**

INTRODUCTION

As we know, news is a vital source for keeping man aware of what is happening around the world. They keep a person well informed and educated. The availability of news on the internet has expanded the reach, availability and ease of reading daily news on the web. News on the web is quickly updated and there is no need to wait for a printed version in the early morning to get updated on the news. The booming news industry (especially online news) has spoilt users for choices. The volume of information available on the internet and corporate intranets continues to increase. So, there is need of text categorization.

Many text categorization approaches exist such as Neural Network, decision tree etc.But the goal of this paper is to use K-mean successfully for text categorization. In the past work has been done on this topic by Susan Dumais, Microsoft Research One, Microsoft Way Redmond and their team. They used inductive learning algorithms like naïve bayes and SVM for text categorization. In 2002, Tan et al. used the bigrams for text categorization successfully. In 2012 Xiqing Zhao and Lijun Wang of HeBei North University, Zhang Jiakou , China improved KNN(K-nearest neighbour) algorithm to categorize data into groups. The current use of TF-IDF (term frequency inverse document frequency) as a vector space model has increased the scope of the text categorization scenario. In this paper, we are using dictionary which gives appropriate result with k-mean. This paper shows the methods which gives the accurate result.

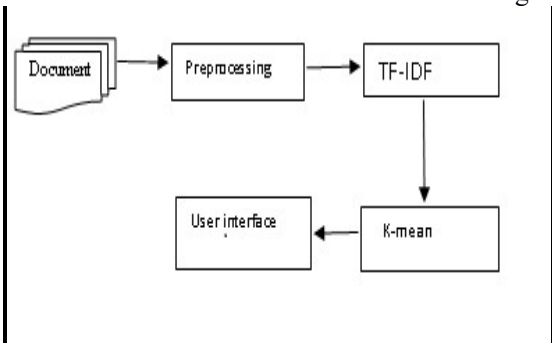
Text categorization is nothing but the process of categorizing text documents on the basis of word, phrases with respect to the categorize which are predefined.

The flow of paper is as follows: - 1st section gives the article categorization, 2nd section gives the application, 3rd gives the conclusion, 4th gives the reference for the paper.

1. Article categorization:-

Article categorization mainly includes five steps. Firstly we have to gather news articles, secondly preprocessing is done, then we have to apply TF-IDF, after that k-mean then user interface is created in which news articles are listed. The following fig. give the steps

of article categorization.



A. Document:-

The news are extracted from various websites like yahoo, Reuters, CNN etc.For our project we use RSS for extraction purpose.RSS is a protocol for providing news summaries in xml format over the internet. This extracted news are stored in HDFS (Hadoop Distribution File System) which is used to stored large amount of data.

B. Preprocessing:-

As large amount of data is stored in HDFS, it is necessary to process the data for fast and accurate result. Some modifications are done on the character such that it should be best suitable for the feature extraction. In this step stop word removal, stemming of data is done.

a. Stop word removal:-

The words like a, an, the, which increases the volume of data but actual of not any use are removed from the data. The process of removing these unused words is known as stop word removal.

b. Stemming:-

Stemming is the process of reducing inflected words to their word stem or root .The root of the word is found by this technnique.eg.The 'running' is reduced to the 'run'.

C.TF-IDF:-

Tf-idf stands for term frequency-inverse document frequency. This term is used to predict the importance of word in specific document. In this technique, the word which is frequently/highly appear in a document have high score .And words which appear in more frequency in many document are having low score. So, we are going to do preprocessing before TF-IDF as it removes the, an, a etc. and also if we ignore the preprocessing then also the words like a, an, the are in many document hence the low score.

Basically, tf-idf is the multiplication of term frequency (tf) and inverse document frequency (idf).Mathematically it is represented as,

$$\text{Tf -idf= tf (t,d)*idf (t,D).}$$

Where,

Tf (t,d)=Raw frequency of term in a document.

Idf (t,D)=It is a frequency/importance of term across all document.

Term frequency:-

Tf (t,d)=0.5+(0.5*f(t,d)/max. frequency of word).

Inverse Term Frequency:-

Idf (t,D)=log[(no .of documents)/(no. of documents in which term t appear)]

D. K-mean:-

K-mean is the simplest algorithm which is used mainly for the clustering. In our project we used mahout as the platform which has its own k-mean driver.

Algorithm:-

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center using:

$$v_i = \frac{x_1+x_2+x_3+x_4+x_5.....x_n}{c_i}$$

Where,

'ci' represents the number of data points in ith cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3.

E .User interface

Lastly, User interface is created which is in XML document. Actually, it is final outcome of the article categorization. Clustering is performed. In simple term, Clustering is nothing but the extraction of meaningful/useful articles from large number of articles.

Result

This project uses Apache based framework support in the form of Hadoop ecosystem, mainly HDFS (Hadoop Distributed File System) and Apache Mahout. We have implemented the project on net beans IDE using Mahout Libraries for k-means and TF-IDF. The code has been deployed on Tomcat server.

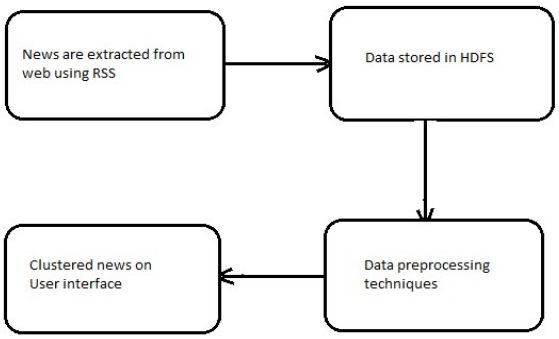


Fig .1. The Flow Diagram of the Project.

As discussed earlier, the database is extracted from the internet using RSS. The data is stored in HDFS. And then stop word removal and stemming is applied to increase the efficiency of data. Then TF-IDF has been used in the project to understand the context of the article by calculating the most occurring as well as the considering the rarest words present in a document. As we know k-mean doesn't worked on the work. It only deals with number. TF-IDF as a vector space model helps convert words to a number format which can be given as input to K-means algorithm in the form of sequence file which has key-value pairs. K-mean algorithm gives us the news in the clustered form.

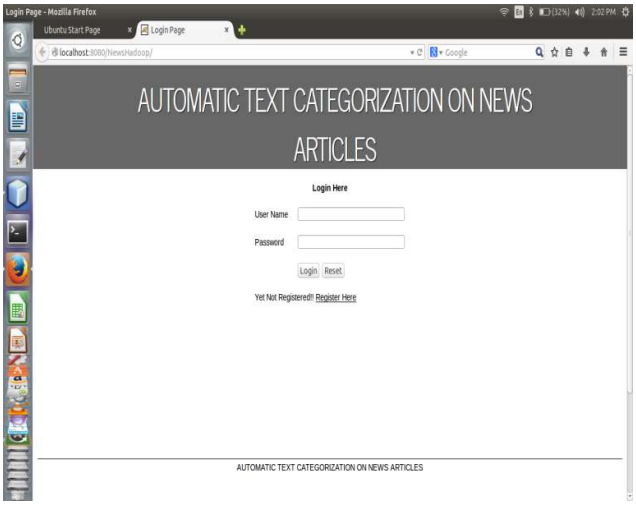


Fig.2. Login Page

The fig. shows the login page. User can login here if he registered himself. If user doesn't have account, then user has to click on the registered here to registered himself.

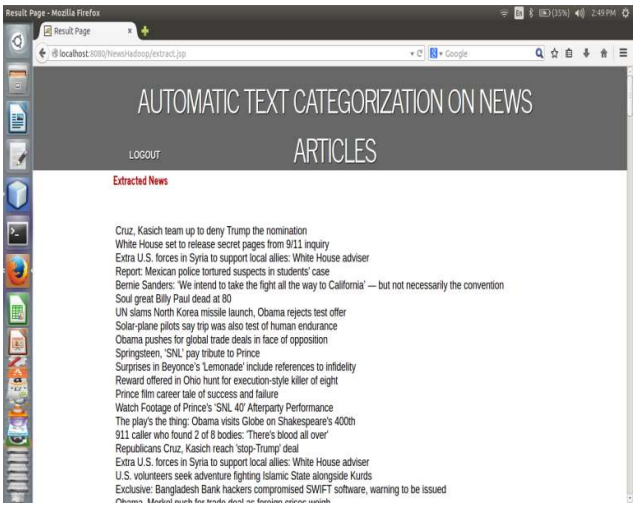


Fig.3.Extracted News

Fig. shows the news extracted from different web which are stored on HDFS

```
204wt: 1.0 distance: 16.768553581114013  vec: Document 0 = [168:5.879,
221:5.879, 335:4.780, 958:5.879, 1123:5.879, 1131:5.474, 1176:5.186,
1267:5.474, 1293:5.879]
246wt: 1.0 distance: 16.5197130328023  vec: Document 1 = [261:5.474,
578:5.186, 794:5.879, 809:5.879, 833:5.879, 955:4.270, 1153:5.879,
1174:5.879, 1354:5.186]
204wt: 1.0 distance: 10.722890881276777  vec: Document 10 = [329:4.780,
464:5.879, 956:4.780, 1247:5.879]
219wt: 1.0 distance: 15.260906660197069  vec: Document 100 = [254:5.474,
263:5.186, 298:5.474, 348:5.879, 753:5.879, 830:5.474, 848:5.879,
1299:3.800]
109wt: 1.0 distance: 19.703869403579272  vec: Document 101 = [70:5.879,
527:5.879, 599:5.474, 728:5.186, 790:5.879, 800:5.474, 837:5.474,
920:5.474, 1102:4.626, 1197:5.879, 1234:5.879, 1296:4.087, 1373:5.879]
211wt: 1.0 distance: 14.9646785895068  vec: Document 102 = [64:4.270,
233:4.963, 428:5.879, 505:4.780, 924:5.879, 1012:5.474, 1018:5.879,
1082:5.186]
198wt: 1.0 distance: 16.099000658443046  vec: Document 103 = [93:4.963,
261:5.474, 283:4.963, 334:4.174, 530:5.186, 700:5.879, 905:5.879,
919:5.879, 1049:5.879]
248wt: 1.0 distance: 16.825784363993815  vec: Document 104 = [34:4.493,
188:5.879, 397:5.879, 610:5.474, 849:5.879, 1098:5.879, 1127:5.879,
1171:5.186, 1270:5.879]
107wt: 1.0 distance: 11.500536245911348  vec: Document 105 = [283:4.963,
556:4.493, 733:5.186, 978:5.186, 1060:5.879]
191wt: 1.0 distance: 15.142052073432632  vec: Document 106 = [165:5.474,
223:4.626, 266:5.186, 574:5.879, 805:4.963, 874:5.879, 1090:5.879,
1289:4.963]
142wt: 1.0 distance: 16.637704019759727  vec: Document 107 = [169:5.879,
265:5.879, 645:5.879, 861:5.879, 1069:5.879, 1164:5.879, 1260:5.879,
1300:5.879]
```

Fig.4.Clustering Process

Above fig.4. Shows the clustering process. The weights are calculated and sequence file is generated using TF-IDF.

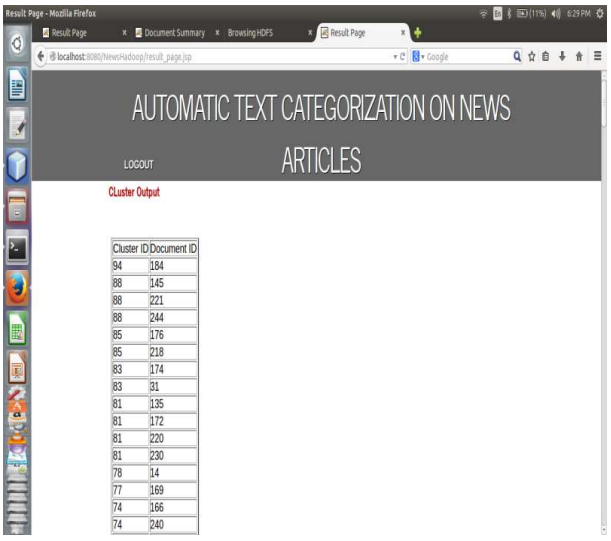


Fig.5.Final Cluster Data

Above fig.5.shows the clustered data which shows the different document with their id in same cluster.

APPLICATION

2.1. Automatic indexing for Boolean information retrieval systems

The application that has stimulated the research in text categorization from its very beginning, back in the '60s, up until the '80s, is that of automatic indexing of scientific articles by means of a controlled dictionary, such as the ACM Classification Scheme, where the categories are the entries of the controlled dictionary. This is typically a multi-label task, since several index terms are usually assigned to each document.

Automatic indexing with controlled dictionaries is closely related to the automated metadata generation task. In digital libraries one is usually interested in tagging documents by metadata that describe them under a variety of aspects (e.g. creation date, document type or format, availability, etc.). Some of these metadata are thematic, i.e. their role is to describe the semantics of the document by means of bibliographic codes, keywords or key-phrases. The generation of these metadata may thus be viewed as a problem of document indexing with controlled dictionary, and thus tackled by means of text categorization techniques. In the case of Web documents, metadata describing them will be needed for the Semantic Web to become a reality, and text categorization techniques applied to Web data may be envisaged as contributing part of the solution to the huge problem of generating the metadata needed by Semantic Web resources.

2.2. Document Organization

Indexing with a controlled vocabulary is an instance of the general problem of document base organization. In general, many other issues pertaining to document organization and filing, be it for purposes of personal organization or structuring of a corporate document base, may be addressed by text categorization techniques. For instance, at the offices of a newspaper, it might be necessary to classify all past articles in order to ease future retrieval in the case of new events related to the ones described by the past articles. Possible categories might be Home News, International, Money, Lifestyles, Fashion, but also finer-grained ones such as The Prime Minister's USA visit.

Another possible application in the same range is the organization of patents into categories for making later access easier, and of patent applications for allowing patent officers to discover possible prior work on the same topic. This application, as all applications having to do with patent data, introduces specific problems, since the description of the allegedly novel technique, which is written by the patent applicant, may intentionally use non standard vocabulary in order to create the impression that the technique is indeed novel. This use of non standard vocabulary may depress the

performance of a text classifier, since the assumption that underlies practically all text categorization work is that training documents and test documents are drawn from the same word distribution.

2.3. Text filtering

Text filtering is the activity of classifying a stream of incoming documents dispatched in an asynchronous way by an information producer to an information consumer. Typical cases of filtering systems are e-mail filters (in which case the producer is actually a multiplicity of producers), newsfeed filters, or filters of unsuitable content. A filtering system should block the delivery of the documents the consumer is likely not interested in. Filtering is a case of binary text categorization, since it involves the classification of incoming documents in two disjoint categories, the relevant and the irrelevant. Additionally, a filtering system may also further classify the documents deemed relevant to the consumer into thematic categories of interest to the user. A filtering system may be installed at the producer end, in which case it must route the documents to the interested consumers only, or at the consumer end, in which case it must block the delivery of documents deemed uninteresting to the consumer.

In information science document filtering has a tradition dating back to the ’60s, when, addressed by systems of various degrees of automation and dealing with the multi-consumer case discussed above, it was called selective dissemination of information or current awareness. The explosion in the availability of digital information has boosted the importance of such systems, which are nowadays being used in diverse contexts such as the creation of personalized Web newspapers, junk e-mail blocking, and Usenet news selection.

2.4. Word Sense Disambiguation

Word sense disambiguation (WSD) is the activity of finding, given the occurrence in a text of an ambiguous (i.e. polysemous or homonymous) word, the sense of this particular word occurrence. For instance, bank may have (at least) two different senses in English, as in the Bank of Scotland (a financial institution) or the bank of river Thames (a hydraulic engineering artifact).

It is thus a WSD task to decide which of the above senses the occurrence of bank in Yesterday I withdrew some money from the bank has. WSD may be seen as a (single-label) TC task once, given a word *w*, we view the contexts of occurrence of *w* as documents and the senses of *w* as categories.

2.5. Hierarchical Categorization of Web Pages

Text categorization has recently aroused a lot of interest for its possible use to automatically classifying Web pages, or sites, under the hierarchical catalogues hosted by popular Internet portals. When Web documents are catalogued in this way, rather than issuing a query to a general-purpose Web search engine a searcher may find it easier to first navigate in the hierarchy of categories and then restrict a search to a particular category of interest. Classifying Web pages automatically has obvious advantages, since the manual categorization of a large enough subset of the Web is not feasible. With respect to previously discussed text categorization applications, automatic Webpage categorization has two essential peculiarities which are namely the hyper textual nature of the documents, and the typically hierarchical structure of the category set.

2.6. Automated Survey coding

Survey coding is the task of assigning a symbolic code from a predefined set of such codes to the answer that a person has given in response to an open-ended question in a questionnaire (aka survey). This task is usually carried out in order to group respondents according to a predefined scheme based on their answers. Survey coding has several applications, especially in the social sciences, where the classification of respondents is functional to the extraction of statistics on political opinions, health and lifestyle habits, customer satisfaction, brand fidelity, and patient satisfaction.

Survey coding is a difficult task, since the code that should be attributed to a respondent based on the answer given is a matter of subjective judgment, and thus requires expertise. The problem can be seen as a single-label text categorization problem, where the answers play the role of the documents, and the codes

that are applicable to the answers returned to a given question play the role of the categories (different questions thus correspond to different text categorization problems).

CONCLUSION

Article categorization is very critical task as every method gives different result. There are many methods like neural network, k-medoid etc but we are using k-mean and TF-IDF, which increases accuracy of the result. Also it cluster the article in proper ways.

In future we will try to solve the article categorization for the data which is obtained by using web crawler thus creating one stop solution for news. Also, sophisticated text classifiers are not yet available which if developed would be useful for several governmental and commercial works. Incremental text classification, multi-topic text classification, discovering the presence and contextual use of newly evolving terms on blogs etc. are some areas where future work can be done.

REFERENCES

[1] Document Clustering, Pankaj Jajoo, IITR.

[2] Noam Slonim and Naftali Tishby. *"The Power of Word Clusters for Text Classification*"School of Computer Science and Engineering and The Interdisciplinary Center for Neural Computation The Hebrew University, Jerusalem 91904, Israel.

[3] Mahout in Action, Sean Owen, Robin Anil, Ted Dunning and Ellen Friedman Manning Publications, 2012 edition

[4] Saleh Alsaleem, Automated Arabic Text Categorization Using SVM and NB,June 2011.

[5] Wen Zang and Xijin Tang,TFIDF , LSI and Multi-word in Information Retrieval and text categorization ,Nov-2008

[6] R.S. Zhou and Z.J. Wang, A Review of a Text Classification Technique: K- Nearest Neighbor, CISIA 2015

[7] FABRIZIO SEBASTIANI, Machine Learning in Automated Text Categorization.

[8] Susan Dumais , David Heckerman and Mehran Sahami, Inductive Learning Algorithms and Representations for Text Categorization

[9] RON BEKKERMAN and JAMES ALLAN, Using Bigrams in Text Categorization,Dec 2013.

[10]Lodhi, J. Shawe-Taylor, N. Cristianini, and C.J.C.H. Watkins. Text classification using string kernels. In Advances in Neural Information Processing Systems (NIPS), pages 563–569, 2000.