I.J. Mathematical Sciences and Computing, **2016, 2, 14-23**
Published Online April 2016 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijmsc.2016.02.02

Available online at ***http://www.mecs-press.net/ijmsc***

Research Domain Selection using Naive Bayes Classification

Selvani Deepthi Kavila [a*], Dr.Radhika Y [b]

[a] *Assistant Professor, Department of CSE, Anil Neerukonda Institute of Technology And Sciences, Visakhapatnam-531162, India.*
[b] *Associate Professor, Department of CSE, Gitam Institute of Technology, Gitam University, Visakhapatnam-530045, India.*

Abstract

Research Domain Selection plays an important role for researchers to identify a particular document based on their discipline or research areas. This paper presents a framework which consists of two phases. In the first phase, a word list is constructed for each area of the research paper. In the second phase, the word list is continuously updated based on the new domains of research documents. Primary area and Sub area of the documents are identified by applying pre-processing and text classification techniques. Naive Bayes classifier is used to find the probability of various areas. An area having the highest probability is considered as primary area of the document. In this paper text classification procedures is condensed as that are utilized to arrange the content archives into predefined classes. Based on the performance analysis, it has been observed that the obtained results are efficient when compared to manual judgement.

Index Terms: Research Domain Selection, Information Retrieval, Text mining, Classification.

1. Introduction

In the internet era, the data and information explosion have resulted in the availability of large volume of data. In the recent past data mining techniques have proved to be successful in gathering knowledge from such large volume of data. Text mining is the task of discovering interesting patterns from large volumes of data, where the data can be stored in data base and in other information repositories. The main idea of text mining process is to extract knowledge from the documents and transform it into an understandable structure for further use. It is a young interdisciplinary field, drawn from areas such as data warehousing, machine learning, data visualization and information retrieval. Other contributing areas include pattern recognition, signal processing, spatial data analysis, neural networks, image databases, and other application fields, such as

* Corresponding author. Tel.: +91-9440187827; fax: 08933-226395
E-mail address: selvanideepthi14@gmail.com

business and economics.Text mining is the process of extracting small pieces of text data from large amounts of unstructured data. The purpose of Text Mining is to process unstructured (textual) information and extract meaningful information from the text.

Now a day there is a tremendous growth in paper publishing in various streams of research. There are many research disciplines and again we have a lot of many sub areas under a particular discipline. By simply looking at the title, a person may not get an idea about which research discipline of the paper. He has to give a thorough reading of the paper and it consumes a lot of time. It would be better if we have some tool which can specify not only the main research area but also the sub area as the researcher may be interested in a specific area under the main domain. This will save time. In this paper, a method is proposed for classifying various papers present in the repository into their respective disciplines and also identifying their sub areas. In the first phase i.e. system training phase lot many papers of various disciplines are collected and stored in the repository. Pre-processing techniques are applied on each paper and the paper is classified into a primary research area by taking the frequency count of index terms into account. In the second phase the sub area is identified with the frequency count of keywords.

Text mining techniques used for classification are applied on static repository of research documents. Now-a-days the number of documents in repository increases with time i.e. new domain of documents related to various areas is added to the repository and it requires identifying and adding feature words in the list. This list is useful at the time of classifying documents of different areas. So there is a need to classify documents using the recent knowledge of the repository after adding new domain of documents. In this paper a framework is proposed that can classify a stream of research documents inwards to the repository. The proposed frame work contains of two stages where it incorporates development of word list for every area of the paper in the first stage. In the second stage it constantly updates the word list associated with the new stream of research documents.

Rest of the paper is described as follows: Section 2 describes the Literature survey based on text mining and classification of the papers using area tracking. Section 3 describes the problem study related to stream of documents dynamically added to repository, the proposed architecture of the system and Methodology are also presented in this section. Section 4 describes the Performance Analysis and Results followed by conclusion and future work in section 5.

2. Literature Survey

2.1. Background Work Related to Text Mining and Classification

Jian ma et al [1] proposed Ontology-Based Text-Mining Method to Cluster proposals for Research Project. For identifying the research areas selection is an important task. When large numbers of research proposals are received, it is necessary to set them according to their similarity in research disciplines. Current methods for grouping papers are based on manual matching of similar keywords. So, in order to overcome the manual matching jian ma et al proposed a system that includes three phases

Phase 1 deal with Constructing Research ontology: The research topics of different disciplines can be expressed by research ontology. Suppose that there are M discipline areas, A_m denotes discipline area m where m=1, 2... M. Research ontology construction is presented in the following steps.

- Step 1: Creating the research topics of the discipline A_m (m = 1, 2, . . . , M). The keywords of the supported research projects are collected each year, and their frequencies are counted. The frequencies of keywords are denoted by the feature set (N_{om}, ID_m, year, {(keyword1, frequency1), (keyword2, frequency2), . . . , (keywordm, frequencym)}), where N_{om} is the sequence number of the m^{th} record and ID_m is the corresponding discipline code.
- Step2: For constructing the research ontology. Firstly, the research ontology is classified on the basis of research areas introduced in the background. It is then developed according to several specific research

areas. Next, it is further divided into some narrower discipline areas.

- Step 3: Update the Research ontology. If the project funding is completed in every year, the research ontology is updated according to agency's policy and the change of the feature set.

Phase 2 Research Proposals are classified into disciplines: Proposals are classified based on the discipline areas to which they belong. An easy sorting algorithm is used for classifying the proposals.

This is illustrated as follows:

- Suppose that there are m discipline areas, and A_m denotes the area m (m = 1, 2. . . M). P_i denotes proposals where i = 1, 2. . . I, and S_m represent the place of proposals which belongs to area m. Then, to classify proposals a sorting algorithm can be implementing to their discipline areas.

Phase 3 deals with Clustering the Proposals Based on Similarities Using Text Mining: After the research proposals are classified by the discipline areas, the proposals in each domain are clustered using the text-mining technique. The clustering process consists of five steps namely text document collection, text document pre-processing, text document encoding, vector dimensional reduction and text vector clustering.

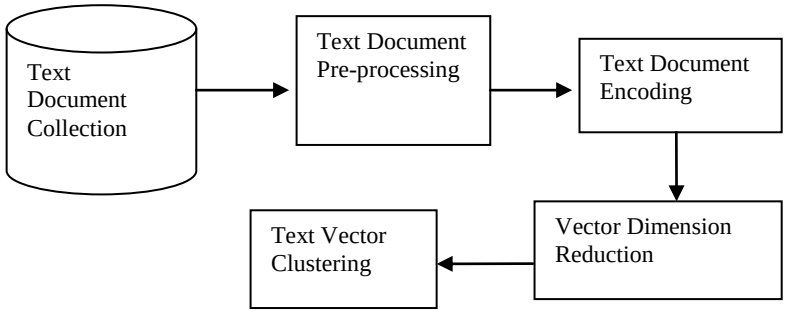

Fig.1. Main Process of Text Mining

Preet kaur et al [2]. Proposed an Ontology based classification and clustering of research proposals and external research reviewers by collecting various research proposals from organizations and then assigning the proposals to the related groups followed by reviewing the proposals. The draw back in the existing system is that the classification of proposals is done manually and also lack of adequate knowledge about the areas. The proposed system uses ontology based text mining method, providing external research reviewers and also provides efficiency for selection of project proposals with increasing number of reviews and research proposals. Ontology is a tree structured knowledge repository. This contains terms and concepts and provides relationship. N.Arunachalam et.al [3] proposed an ontology based text mining frame work for R&D project selection. The phases of this system are constructing new research ontology, classifying new research proposals using Topic identification algorithm, clustering the proposals in to their respective disciplines by using self organized mapping and finally balancing research proposals and regrouping them by considering the applicants.

Topic identification approach:

- Split the text into sentences by using text splitter tool.
- Parse the sentences i.e., find candidates and eliminate useless calculation.
- Select the candidate parts
- Compute the weight for each candidate topic
- Select the final topic.

Steps of Classification Process:

- Data pre-processing
- Defining the training set and test sets
- The classification algorithm is created selectively by using the classification model
- Classification model validation
- Classification of unknown/new text documents.

Chen [4] proposed a fuzzy-logic-based model as a tool for document selection. Henriksen et al [5] presented a rank based tool for document verification and identification. Ghasemzadeh et al [6] explained a decision support frame work for identifying the area. Machacha et al [7] proposed a fuzzy logic based frame work to identify the area. Butler [8] explained a theory for research document ranking and selection. Loch et al. [9] established an optimized model for document selection, while Murad et al [10] developed an analytic network process model. Greiner [11] explained a frame work to support project selection, and Tian suggested an approach for selecting R&D documents. Choi et al. [12] explained text-mining approach for document extraction and filtering. Sun [13] used hybrid knowledge based model document identification. Hettich explained [14] a text-mining approach to combine documents and assigned an area to the document. There are many text-mining works that can be used to classify documents [15][16], which are developed with a focus on English text.

2.2. Text Classification

Text (or Document) classification is an active research area of text mining, where the documents are classified into predefined classes. Most of the text documents include letters, newspapers, articles, blogs, technical reports, proceedings, and journal papers, etc. Document Filtering is also based on the classification algorithm to extract the relevant documents related to specific topic from the set of documents.

Text Classification tasks can be divided into two types: supervised document classification has peripheral mechanism such as human criticism gives data on the suitable classification for documents and the other is unsupervised document classification otherwise called document clustering, where the classification should be done with no outer reference, this framework does not have predefined classes. There is other task called semi supervised document classification, where some documents are named by the external mechanism i.e. some documents are already classified for better learning of the classifier.

- Need for Automatic Text Classification: Manual classification of millions of text documents is an expensive and time taking task. Hence, automatic text classifier is constructed using pre-classified sample documents whose accuracy and time efficiency is much better than manual text classification. In this paper we summarize text classification techniques that are utilized to characterize the content archives into predefined classes.

2.3. Text Pre-Processing

Pre-processing phase is used to determine the most important keywords that are meaningful, leaving behind the keywords that do not contribute to the differentiation between the documents. The pre-processing task includes the processing of textual data in a structure of data-mining-ready, where the most important text features that serve to differentiate between text categories are found. It is the process of adding a new document into an information retrieval system. An effective pre-processor represents the document efficiently in terms of both space (for storing the document) and time (for processing retrieval requests) requirements and maintain good retrieval performance (precision and recall).

2.4. Topic Identification

Topic is a stream of words which stands for the content of a text. Knowing the topic of a document can help the people to be aware of its content thereby facilitating their searching process. There is a difference between 'the topic' and 'the title'. Title is also a sequence of terms but rather represents the name of a study and does not necessarily represent the content of this study. However, the title does not necessarily stand for the content of documents and it is not possible to judge about the content of documents by their titles only. The automatic identification of the topic of a given document is not an easy task as a document may contain multiple topics.

3. Proposed System

There are two phases in classification. In training phase it reads the text from document line by line, finds the count of feature words and compares with the words stored in the feature vector file along with label. In this process it calculates the frequency of each feature word. In testing phase it finds the counts of feature words to classify the documents.

String matching algorithm is used to find all occurrences of a pattern in a given text. i.e. p[1...m] in text T[1..n] where n>=m .

Step 1: Constructing research knowledge based system

A repository is maintained which consists of various domain papers and also datasets which consists of keywords of various domains. After that Pre-processing techniques are applied to the documents. Here stop words are removed from the documents, then pattern matching is performed with the extracted words from the documents with the data set keywords.

Step 2: Classifying the papers into respective disciplines

For classifying the documents, Naive Bayes classification is used to find the probability of various areas. Among various areas the highest probability value is taken as the primary area of the paper. The same procedure is applied to find the sub area.

- L: Set of feature words of order d_{xm}, doc_t: input document to check, number of classes: d i.e. C_1, C_2, $C_3 \ldots C_d$.
- Naïve Bayes classifier predicts X belonging to Class C_i
- if $P(C_i/X) > P(C_j/X)$ for $1<= j <= m, j <> i$
- $P(C_i/X) = P(X/C_i)\, P(C_i) / P(X)$
- Maximize $P(X/C_i)\, P(C_i)$ as $P(X)$ is constant
- Where $c_1, c_2, c_3 .. c_n$ are the data sets of various domains like data mining, image processing, networks, and network security. And 'x' is the list of words in the documents.

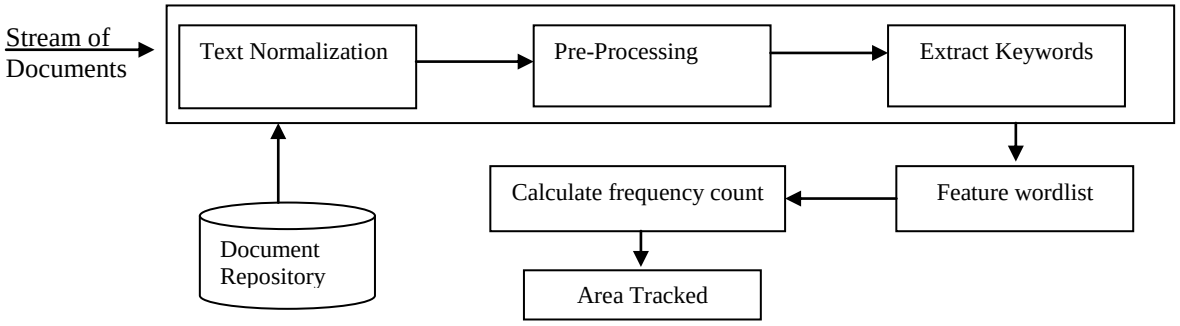

Fig.2. System Architecture

4. Performance Analysis and Results

It is important to measure the performance for an information retrieval system. In this section, some of the common measures that have been used in the paper are described. To evaluate the experimental results, several standard measures such as precision and recall are used. The precision is the ability to retrieve the top most documents that are mostly relevant. The recall is the ability of the search to find all of the relevant items in the repository. The F-measure is a weighted mean of precision and recall, it is also known as balanced F-score.

$$\text{Precision (P)} = \frac{\text{No. of Relevant documents Retrieved}}{\text{Total No. of documents Retrieved}}$$

$$\text{Recall(R)} = \frac{\text{No. of Relevant documents Retrieved}}{\text{Total No of relevant documents}}$$

$$\text{F} - \text{measure} = \frac{2\text{PR}}{\text{P} + \text{R}}$$

The documents are partitioned into four sets namely relevant or not relevant and retrieved or not retrieved. It is aimed to find the efficiency of the system. By calculating the precision, recall and f-measure the results of manual judgment are compared with the tool which was trained to the system.

Table 1. Relevant and Retrieved Sets

	Relevant	Not relevant
Retrieved	T_p	F_p
Not retrieved	F_n	T_n

Where, T_p and F_p are true positive and false positive respectively; F_n and T_n are false negative true negative respectively.

In the repository, 1000 documents are maintained and for these documents the primary area and sub area of the document are identified. Here manual judgment results are compared with the system generated results.

A manual judgement is generated for all the research papers. The primary and secondary areas of particular papers are noted. A sample judgement file is shown below for the first 10 papers that are studied.

Table 2. Sample Judgement Documents

Document Number	Primary Research Area
D1	Data Mining
D2	Image Processing
D3	Data Mining
D4	Image Processing
D5	Data Mining
D6	Data Mining
D7	Data Mining
D8	Image Processing
D9	Networks
D10	Data Mining

The repository consists of papers belonging to various areas. Primary and Subareas are identified for each paper. The different areas maintained here are data mining, networks, image processing, and network security.

The below Table 3 shows the performance of this system for the first 10 papers and similar observations are made for the rest of the papers.

Table 3. Compare Manual Judgement with System Generated

Document Number	Primary Research Area as per Manual Judgment file	System Generated Research Area
D1	Data Mining	Data Mining
D2	Image Processing	Data Mining
D3	Data Mining	Data Mining
D4	Image Processing	Image Processing
D5	Data Mining	Data Mining
D6	Data Mining	Networks
D7	Data Mining	Data Mining
D8	Image Processing	Image Processing
D9	Networks	Networks
D10	Data Mining	Data Mining

The following table generates the calculation of performance measures for all the 1000 papers:

Table 4. Domains and retrieved documents

Domain	Retrieved	Not retrieved	Not relevant	Not relevant not retrieved
Data Mining	215	115	90	580
Networks	112	21	17	850
Image processing	240	29	11	720
Network security	127	14	9	850

The values of precision, recall and f-measure are calculated as,

For Data mining, P= 215/ (215+90) = 0.7049;
R= 215/ (215+115) =0.6575
For Networks, P= 112/ (112+17) = 0.8682;
R=112/ (112+21) = 0.8421
For Image processing, P = 240/ (240+11) =0.9561;
R = 240/ (240+29) = 0.8921
For Network security, P= 127/ (127+9) = 0.9338;
R= 127/ (127+14) = 0.9007

The f-measure of Data mining = (2*0.7049*0.6575)/ (0.7049+0.6575) = 0.6803
The f-measure of Networks = (2*0.8682*0.8421)/ (0.8682+0.8421) = 0.8549
The f-measure of Image processing = (2*0.9561*0.8921)/ (0.9561+0.8921) = 0.9229
The f-measure of Network security = (2*0.9338*0.9007)/ (0.9338+0.9007) = 0.9169

Table 5. Calculation of Precision, Recall, F- measure

Domain	Precision	Recall	F-measure
Data mining	0.7049	0.6575	0.6803
Networks	0.8682	0.8421	0.8549
Image processing	0.9561	0.8921	0.9229
Network security	0.9338	0.9007	0.9169

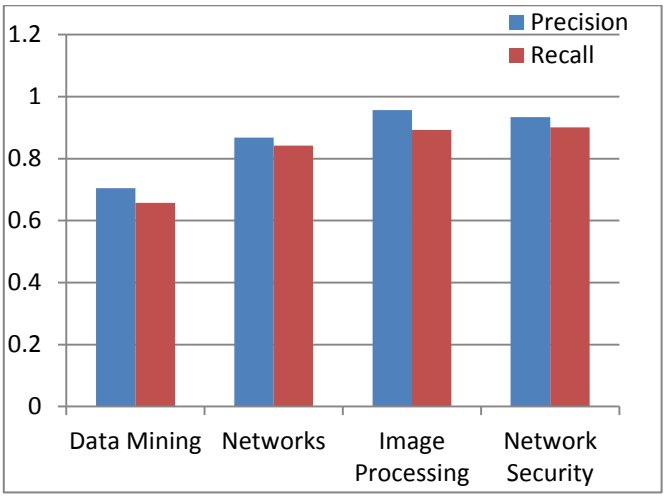

Fig.3. Performance Evaluation

Fig. 3 depicts the performance of the system for all the different classes of papers. When compared with manual evaluation where a list is maintained for each class of paper, our results are promising. The x-axis represents the domains of the documents like data mining, networks, image processing, network security and the y-axis represents the scale of intervals. If precision reaches 1 it shows that the system is more efficient**.**

5. Conclusion and Future Work

In this paper a framework is proposed which facilitates a researcher to recognize a document based on its domain. This technique is capable of identifying the main research area as well as the sub area by creating a word list for each area of paper and updating it as new domains are added. The area of a document is identified by constructing a repository initially and classifying the new proposals using the topic identification algorithm which are then clustered into their respective disciplines. Standard measures such as precision and recall are used for the evaluation of the system. A comparison is made between the system yielded results and the human produced ones to determine the efficiency. In future the work will be expanded to larger datasets and apply other classification techniques to compare efficiencies. Also one may focus on finding a match between scientific papers and the external reviewers. In future focus may be levied on to find a systematic way of assigning papers to reviewer.

References

[1] Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, Ou Liu .An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection. IEEE transactions on systems, man and cybernetics part a: systems and humans, vol. 42, no 3, may 2012.

[2] Richa Sapra and Preet Kaur. Ontology Based Classification And Clustering Of Research Proposals and External Research Reviewers. International Journal of Computers & Technology, Volume 5, No. 1, May - June, 2013.

[3] N.Arunachalam, S.Hismath Begum, E.Sathya and M.Uma Makeswari. An Ontology Based Text Mining Framework for R&D Project Selection. International Journal of Computers and Technology, volume 5, No.1, February 2013.

[4] N. Gorla and K. Chen. Information system project selection using fuzzy logic. IEEE Transactions and Systems, Man and Cybernetics society, Systems and Humans, vol. 28, no. 6, pp. 849–855, Nov. 1998.

[5] A. D. Henriksen and A. J. Traynor. A practical R&D project-selection scoring tool. IEEE Transactions Engineering Management. vol. 46, no. 2, pp. 158–170, May 1999.

[6] F. Ghasemzadeh and N. P. Archer. Project portfolio selection through decision support. Decision Support Systems, vol. 29, no. 1, pp. 73–88, Jul. 2000.

[7] L. L. Machacha and P. Bhattacharya. A fuzzy-logic-based approach to project selection. IEEE Transactions Engineering Management. vol. 47, no. 1, pp. 65–73, Feb. 2000.

[8] J. Butler, D. J. Morrice, and P. W. Mullarkey. A multiple attribute utility theory approach to ranking and selection. Management Sciences, vol. 47, no. 6, pp. 800–816, Jun. 2001.

[9] C. H. Loch and S. Kavadias. Dynamic portfolio selection of NPD programs using marginal returns. Management Sciences, vol. 48, no. 10, pp. 1227– 1241, Oct. 2002.

[10] Murad Habib, Raza Khan and Javaid L. Piracha. Analytic network process applied to Research & Development project selection.

[11] M. A. Greiner, J. W. Fowler, D. L. Shunk, W. M. Carlyle, and R. T. Mcnett.A hybrid approach using the analytic hierarchy process and integer programming to screen weapon systems projects. IEEE Transactions Engineering and Management. vol. 50, no. 2, pp. 192–203, May 2003.

[12] C. Choi and Y. Park.R&D proposal screening system based on text mining approach.Int. J. Technology Intelligence Plan., volume. 2, no. 1, pp. 61–72, 2006.

[13] Y. H. Sun, J. Ma, Z. P. Fan, and J. Wang. A hybrid knowledge and model approach for reviewer assignment. Expert Systems. Applications, volume 34, no. 2, pp. 817–824, Feb. 2008.

[14] S. Hettich and M. Pazzani.Mining for proposal reviewers: Lessons learned at the National Science Foundation. in Proceedings. 12th International Conference Knowledge Discovery. Data Mining, 2006, pp. 862–871.

[15] C. P. Wei and Y. H. Chang.Discovering event evolution patterns from document sequences.IEEE Transactions Systems, Man, Cybern. A, Systems, Humans, vol. 37, no. 2, pp. 273–283, Mar. 2007.

[16] T. H. Cheng and C. P. Wei.A clustering-based approach for integrating document-category hierarchies. IEEE Transactions Systems, Man, Cybern. A, Systems, Humans, vol. 38, no. 2, pp. 410–424, Mar. 2008.

Authors' Profiles

K.Selvani Deepthi is currently Pursuing PhD from Gitam University of Visakhapatnam. She is working as Assistant Professor in Anil Neerukonda Institute of Technology and Sciences. Her area of Interest is Natural Language Processing, Text Mining and Data Mining.

Y.Radhika is doctorate in computer science and engineering. She is working as Associate Professor in Gitam Institute of Technology, Gitam University. Her area of Interest is Data Mining.

How to cite this paper: Selvani Deepthi Kavila, Radhika Y,"Research Domain Selection using Naive Bayes Classification", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.2, No.2, pp.14-23, 2016.DOI: 10.5815/ijmsc.2016.02.02