ISSN: 0976-3104
SPECIAL ISSUE
Danti and Bhushan.

ARTICLE OPEN ACCESS

CLASSIFICATION OF TEXT DOCUMENTS USING INTEGER REPRESENTATION AND REGRESSION: AN INTEGRATED APPROACH

Ajit Danti[1] and SN Bharath Bhushan[2*]
[1]*Department of Computer Science, College of Computer Science, King Khalid University, SAUDI ARABIA*
[2]*Karnataka Government Research Center, Sahyadri College of Engineering and Management, Mangalore, KA, INDIA*

ABSTRACT

Text Classification approaches is receiving more and more attention due to the exponential growth of the electronic media. Text representation and classification issues are usually treated as independent problems, but this paper illustrates combined approaches for text classification system. Integer Representation is achieved using ASCII values of the each integer and later linear regression is applied for efficient classification of text documents. An extensive experimentation on four publically available corpuses are carried out to show the effectiveness of the proposed model which provide better result as compared to static fault models. This paper provides a study and detailed analysis of result and experiment performed.

Received on: 18th-June-2015
Revised on: 20th-July-2015
Accepted on: 03rd– August-2015
Published on: 8th-Jan-2016

KEY WORDS

Integer Representation, Text Classification, Regression based searching.

***Corresponding author: Email:** ajitdanti@yahoo.com, sn.bharath@gmail.com; **Tel:** +91-9480766063

INTRODUCTION

Text data, especially the increased popularity of the Internet and the World Wide Web became the most common types of information to store house. Most common sources are web pages, emails, newsgroup messages, internet news feeds etc,[1]. Many real time text mining applications have gained a lot of attention due to large production of textual data. Many applications of text classification are spam filtering, document retrieval, routing, filtering, directory maintenance, ontology mapping, etc.

The goal of the text classification algorithm is to identify text documents with the ontology of domains defined by the subject experts. In text classification a boolean value will be assigned to each pair where is a set of predefined categories and is the theme of documents. The task is to approximate the true function by means of a function such that coincide as much as possible. The function is called a classifier. A classifier can be built by training it systematically using a set of training documents [1]. Generally, textual data being unstructured in nature, pose a number of challenges such as desired representation model, high dimensionality, semanticity, volume and sparsity. Some solutions for these challenges can be found in [2].

In this paper an integer representation for text document which minimize the amount of memory required to store a word which in turn reduces the processing cost is proposed. Text representation algorithm works on the principle that, an integer number requires minimum memory when compared to store a word. Integer representation based classification of text documents is an unconventional approach for classification of text documents.

The rest of the paper is organized as follows. In section 2 a brief literature survey on the text classification is presented. In section 3, a proposed model for the compression based classification of text document. Section 4 discusses about experimentation and comparative analysis performed on the proposed models. Paper will be concluded in section 5.

THE IIOAB JOURNAL
ISSN: 0976-3104

LITERATURE SURVEY

In literature few works of compression based text classification can be seen. Generally, text compression involves context modeling which assigns a probability value to new data based on the frequencies of the data that appeared before. But, context modeling algorithms suffers from slow running time and require large amount of main memory. Mortan [3] proposed a modeling based method for classification of text documents. A cross entropy based approach for text categorization is presented in [4]. It is based on the fact that the entropy is the measure of information content. Compression models are derived from information theory, based on this theoretical fact, language models are constructed for text categorization problem. Authors have also illustrated that, character based prediction by partial matching (PPM) compression schemes have considerable advantages over word based approaches. Frank [5] considered the task of text classification as a two class problem. Different language models such as Ma and Mb are constructed for each class using PPM methods. Test document will be compressed according to different models and gain per model is calculated. Finally class label will be assigned based on the positive and negative gain. Modeling based compression for low complexity devices is presented in [6]. This method is based on the fact that, PPM based approaches require high computational effort which is not practically advisable for low complexity devices such as mobile phones. Algorithm makes use of static context models which efficiently reduce storage space. Similar type of work for low complexity devices are found in [7]. This approach split the data into 16 bit followed by the application of Quine-McCluskey Boolean minimization function to find the minimized expression. Further static Huffman encoding method is used for text compression. Dvorski proposed an indexed based compression scheme for text retrieval systems. A document is considered as a combination of words and non words. Similarly, Khurana and Koul [8] considered English text as a dictionary where each word is identified by a unique number in which a novel algorithm is proposed which consists of four phases, where each phase is for different type of input conditions along with a technique to search for a word in the compressed dictionary based on the index value of the word. Word based semi static huffman compression technique is presented [9] in which algorithm captures the words features to construct byte oriented huffman tree. It is based on the fact that byte processing is faster than bit processing. Automaton is constructed based on the length of search space. End tagged Dense code compression is proposed in [10]. Though the proposed method looks similar to tagged huffman technique, the algorithm has the capacity in producing better compression ratio, constructing a simple vocabulary representation in less computation time. Compressed strings are 8% shorter than tagged huffman and 3% over conventional huffman. Another word based compression approach is found in [11]. This method compresses the text data in two levels. First level is the reduction which is through word look up table. Since word look up table is operated by operating system, the reduction is done by operation system only. According to this method each word will be replaced by an address index. Next stage is the compression stage. Deflatge compression algorithm is used for compression. Four different huffman, W-LZW, word based first order and first order context modeling methods for text compression are presented in [12]. All the word based compression techniques discussed above, maintain two different frequency tables for words and non- words.

Many approaches for classification of text documents can be found in literature. These approaches include naïve bayes [13, 14], nearest neighbor [15-17], decision trees [18], support vector machines [19] and neural network [20] approaches.

PROPOSED METHOD

In this paper, proposed model can be categorized into two stages, such as text representation stage and regression based searching stage.

Text Representation

It is theoretically verified that a sequence of characters requires more memory than an integer number. Based on this, a novel text representation algorithm is proposed, which has the facility of representing character string (a word) by a unique integer number. It is known that, words are the collection of alphabets, which represent a specific meaning. Similarly a text document is collection of such strings which represent a specific domain. Words from the text documents are extracted. Each word is then subjected for compression algorithm and then it is represented by an integer number. The whole procedure is algorithmically represented in the algorithm 1 and pictorially represented in **Figure– 1** and the proposed method is explained in illustration-2.

THE IIOAB JOURNAL
ISSN: 0976-3104

Cumulative sum of ASCII value is determined for given textual word using the equation (1)

$$Cw = \sum_{k=1}^{n} a_k \, b^k \quad \ldots(1)$$

Where,

C_w = Cumulative sum of ASCII values

w = length of the document. (Number of words in the documents).

k = number of alphabets in the word.

a = ASCII value of alphabet.

b = base.

Illustration : 1

Input word: **heart**.

ASCII values: 104,101,97,114,116.

$= 104 \times 2^4 + 101 \times 2^3 + 97 \times 2^2 + 114 \times 2^1 + 116 \times 2^0 =$ **3204.**

Data	Before Compression	After Compression
heart	5 bytes	2 bytes

heart will be represented by an integer number **3204**.

Fig: 1. Pictorial representation of Compression algorithm

Regression Based Searching Stage

Now, the text document can be viewed as a collection of integer vales. As a result text classification problem got reduced into integer searching problem. Once the data is represented by an integer value, it is sorted and linear regression is applied using the equation 2.

$$a_0 = \frac{\sum x_i y_i \sum x_i - \sum x_i^2 \sum y_i}{(\sum x_i^2 - n\sum x_i^2)}$$

$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (n\sum x_i^2)} \quad \ldots(2)$$

Where, x = Positional Value, and y = word.

Once the regression algorithm is applied, the data will represented by a straight line, as shown in **Figure– 2** using the equation (3)

$$y = a_1 x + a_0. \quad \ldots(3)$$

where *x* gives the appropriate position of the search key element as shown in **Figure– 2**.

THE IIOAB JOURNAL
ISSN: 0976-3104

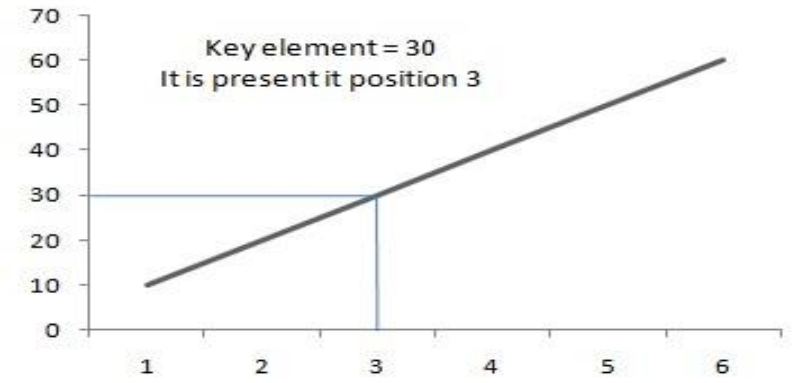

Fig: 2. Regression based searching

Classification of text documents will be accomplished by subjection training documents to representation algorithm. Once the data representation stage is accomplished linear regression is calculated for each class. Effect of this process, training phase will be seen as collection linear regression values. Then, test documents will be subjected to the compression algorithm. As a result, test documents will be collection of integer number. Each integer number from testing document is considered and is fed to regression based searching algorithm. The main advantage of the regression is that, it takes minimum computational unit to search an integer number from the database. Similarly the procedure is carried out and class label will be given to all the integer values. Test document will be assigned a class label based on the maximum class label assigned each integer. **Figure– 3** present the block diagram of the proposed method.

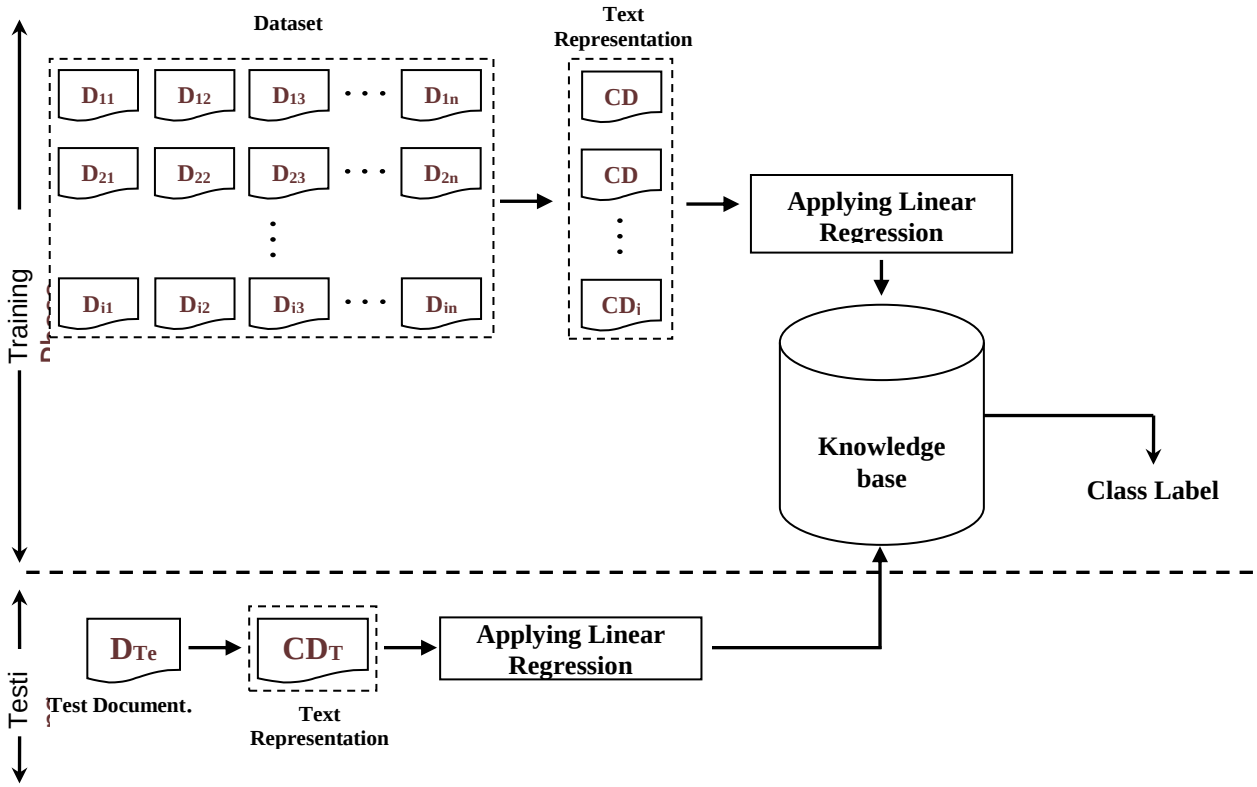

Fig: 3. Block diagram of the proposed approach

THE IIOAB JOURNAL
ISSN: 0976-3104

EXPERIMENTATION

The proposed method is evaluated for its effectiveness and efficiency on the publicly available datasets. The first data is from Wikipedia pages which include the characteristics of the vehicle that the vehicle is in the data. The second dataset includes 10 different classes for 1000 documents are from Google newsgroup data. The third and fourth is from 20 Mini newsgroup and 20 newsgroup datasets. To check the efficiency of the proposed model, two sets of experiments are conducted. First set of experimentation consists of 40% training and 60% testing on all the four datasets. Second set of experimentation consists of 60% training and 40% testing on all the four datasets. The details of the first and second set of experiments are shown in **Table– 1**.

Table: 1. Classification result of proposed compression and regression based technique

Datasets	40% : 60%				60% : 40%			
	Num of Training	Num of Testing	f Measure	Class Accuracy	Num of Training	Num of Testing	f Measure	Class Accuracy
Vehicle Wikipedia	44	66	0.9273	92.61	66	44	0.9476	94.70
Google Newsgroup	40	60	0.9205	92.00	60	40	0.9321	93.16
20 Mini Newsgroup	40	60	0.9003	90.00	60	40	0.9184	92.25
20 Newsgroup	400	600	0.8873	88.62	600	400	0.8960	89.47

F measure is considered for the evaluation of the proposed methods, precision, recall and class accuracy (CA) for each set of experiments using the equation (4),(5),(6) and (7). Let a,b,c and d respectively denote the number of correct positives, false negatives, false positives and correct negatives.

$$fMeasure = \frac{2PR}{P+R} \quad \ldots (4)$$

Where,

P(Precession) = a / (a + c) . . . (5)
R(Recall) = a /(a+d) . . . (6)
CA(Class Accuracy) = (a + d) / N . . . (7)

CONCLUSION

This paper illustrates a method of providing an integer representation of text and regression for classification of text documents is presented. An extensive experimentation is carried on four publically available datasets to show the efficiency of the proposed models. The performance evaluation of the proposed method is carried out by performance measures such as f-measure and class accuracy (CA). The proposed model is very simple and computationally less expensive. One can think of exploring the proposed model further for other applications of text mining. This can be one of the potential directions which might unfold new problems.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

FINANCIAL DISCLOSURE

No financial support received to carry out this research.

THE IIOAB JOURNAL
ISSN: 0976-3104

REFERENCES

[1] Rigutini L. [2004] Automatic Text Processing: Machine Learning Techniques.Ph.D. Thesis, University of Siena.
[2] Sebastiani F.[2002] Machine learning in automated text categorization. *ACM Computing Surveys* 34:1 – 47.
[3] Mortan Y, Nign Wu, and Lisa Hellerstein. [2005] On compression based text classification. Advances in information retrieval in Advances in information retrieval, pages 300–314.
[4] Teahan W, and Harper D. [1998] Using compression based language models for text categorization. Proceedings of 2001 workshop on language modeling and information retrieval.
[5] Frank E, Cai C, Witten H. [2000] Text Categorization using compression models. In proceedings of DCC-00, *IEEE Data compression conferenc*e.
[6] Clemens S and Frank P. [2006] Low complexity compression of short messages. *In proceedings of IEEE Data Compression Conference*, 123–132.
[7] Snel V, Plato J., and Qawasmeh E. [2008] Compression of small text files. *Journal of Advanced Engineering Informatics Information Achieve*, 20: 410–417.
[8] Khurana U and Koul A. [2005] Text compression and superfast searching. Proceedings of the CoRR, 2005.
[9] Moura E, Ziviani N and Navarro and Yates RB. [1998] Fast searching on compressed text allowing errors. Proceedings of the 21st annual international ACM Sigir conference on Research and Development in Information retrieval, pages 298– 306.
[10] Nieves G, Brisaboa, Eva L, and Param J.[2003] An efficient compression code for text databases. Proceedings of the 25th European conference on IR research, pages 468–481.
[11] Azad AK, Ahmad S, Sharmeen R, Kamruzzana SM. [2005] An efficient technique for text compression. In proceedings of *International Conference on Information Management and Business*, 467–473.
[12] Horspool RN and Cormack GV. [1992] Constructing word based text compression of short messages. *In Proceedings of the IEEE Data compression conference*, 62–71.
[13] Hava O, Skrbek M, and Kordík P.[2013] Supervised two-step feature extraction for structured representation of text data. *Journal of Simulation Modelling Practice and Theory*, 33: 132–143.
[14] Rocha L, Mourao F, Mota H, T.Salles, MA Gonc-alves, and W.Meira. [2013] Temporal contexts:Effective text classification in evolving document collections. *Journal of Information Systems*, 38: 388–409.
[15] Meiling Wu, Shengyi J, Guansong and Limin Kuang. [2012] An improved k-nearest neighbour algorithm for text categorization. Expert Systems with Applications, 39:1503–1509.
[16] Zhao Y and Wang Y. [2012] Text Categorization Based on Emergency domain Words : A System Engineering View. *Journal of Systems Engineering Procedia*, 5: 8–14.
[17] Ajit Danti and SN Bharath Bhushan. [2013] Document Vector Space Representation Model for Automatic Text Classification. *In Proceedings of International Conference on Multimedia Processing, Communication and Information Technology*, Shimoga. pp. 338–344.
[18] Lewis DD and M Ringuette. [1998] A comparison of two learning algorithms for text classification. Proceedings of the 3rd Annual symposium on Document Analysis and Information Retrieval, pp. 81 – 93.
[19] Wanga S, D Li, L.Zhao and J Zhang. [2013] Sample cutting method for imbalanced text sentiment classification based on BRC. *Journal of Knowledge-Based System* 37: 451–461.
[20] Patra A, Singh D. [2013] Neural Network Approach for Text Classification using Relevance Factor as Term Weighing Method. *International Journal of Computer Applications*, 68(17): 37-41.

ABOUT AUTHORS

Prof. Ajit Danti *is currently working as professor in Kingdom of Saudi Arabia. He is currently faculty of Department of Computer Science, College of Computer Science, King Khalid University; He received his PhD degree from Gulbarga University, Gulbarga, India. His research interest includes Digital Image Processing, Face Recognition.*

Mr. SN Bharath Bhushan *is currently working as assistant professor in Sahyadri Engineering of College and Management Adyar, Mangaluru, Karnataka, India. He is currently faculty in the department of Computer Applications. He received his MS degree from University of Mysore, Mysore, India. He is pursuing PhD in Text Mining. His research interest includes Text Data Mining and Image processing.*