Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 314 – 320

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

A Lexical Approach for Text Categorization of Medical Documents

Rajni Jindal[a], Shweta Taneja[b,*]

[a] Professor, Dean, Research and Collaboration Head, Department of IT, IGDTUW Delhi-110006 India.
[b]Research Scholar, Dept. of Computer Engineering, DTU Bawana Road, Delhi-110042. India.

Abstract

This research proposes a novel lexical approach to text categorization in the bio-medical domain. We have proposed LKNN (Lexical KNN) algorithm, in which lexemes (tokens) are used to represent the medical documents. These tokens are used to classify the abstracts by matching them with the standard list of keywords specified as MESH (Medical Subject Headings). It automatically classifies journal articles of medical domain into specific categories. We have used the collection of medical documents, called Ohsumed, as the test data for evaluating the proposed approach. The results show that LKNN outperforms the traditional KNN algorithm in terms of standard F-measure.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

*Keywords:*Text Categorization; Lexical Analysis; Tokens; KNN

1. Introduction

In today's world, the role of text categorization is increasing in the field of information retrieval due to the ease of availability of information in digitized form. The process of text categorization[8] is to assign a document into an appropriate category in a predefined set of categories. Earlier, this process had been performed manually.

* Corresponding author. Tel.: +91-921-238-8121; fax: +0-000-000-0000 .
*E-mail address:*shweta_taneja08@yahoo.co.in

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.026

But, due to the increase in the number of documents exponentially, the work cannot be performed manually as it requires a huge amount of time and cost. For example, there is a dataset of medical journal articles, MEDLINE corpus, it requires a considerable effort to perform categorization using Medical Subject Headings (MeSH) categories[1, 19]. This has resulted in a number of researches for automatic text categorization techniques including the Bayesian classifier[2] , the decision tree[3] , the k-nearest neighbor classifier (k-NN)[4] , the rule learning algorithm[5] , neural networks[6] , the fuzzy logic-based algorithm[7] and SVM (support vector machine) [8].

In the field of medical documents, the concepts and techniques of text categorization is being used nowadays. We also know that the text documents suffer from the problem of Curse of Dimensionality[13] . A number of methods are suggested for reducing the dimensionality of text documents. Some of them are: nonlinear dimension reduction techniques [14], discretizing high-dimensional data[15], latent semantic indexing (LSI)[21] and Document Frequency (DF)[16] etc. We have proposed a lexical approach, where we identify tokens or lexemes in the document. Each individual document is represented as a vector of tokens. Our proposed approach serves dual purpose: firstly it reduces the size of the document and further it helps in categorization of documents. We have used KNN algorithm for text categorization. It is a novel concept which has been evaluated empirically on Ohsumed test data collection. Our proposed approach outperforms the traditional KNN algorithm.

The structure of the paper is as follows: Section 2 discusses the background work done in this field. Section 3 describes the proposed LKNN algorithm with its architecture. Section 4 explains the results obtained on Ohsumed collection. Section 5 shows the comparison between traditional KNN and LKNN. Finally, Section 6 concludes the contributions followed by references.

2. Background Work

A lot of work has been done in the field of Text Categorization of medical documents. Authors have proposed different methods for categorizing the documents.

In[9] , texts have been encoded into tables and then similarity measure between tables is calculated and applied to categorization of the bio-medical texts. In another work[10], a new method using rule-based approach was proposed for text categorization. In that method, authors introduced the idea of lexical syntactic patterns as classification features. A novel framework ROLEX-SP was proposed to solves the problem of text categorization. In another work[17], principal component analysis (PCA) method has been used in the field of text mining. They focused on two of its variants namely the neural PCA and kernel PCA for categorization of text documents by extracting semantic concepts. In[11] , authors evaluate and study some machine learning methods: k nearest neighbor (KNN), support vector machines (SVM), naive Bayes (NB) and clonal selection algorithm based on antibody density (CSABAD). According to their concept, only those cells that have higher similarity and lower density are selected to grow. They have proved that SVM and CSABAD perform extensively better than KNN and naive Bayes.

In our work, we have used a novel lexical approach. It is based on identifying tokens or lexemes in the document. And further KNN algorithm is used for text categorization. We have taken KNN algorithm as it is the most popular algorithm and an efficient one.

3. Proposed Lexical KNN Approach (LKNN)

We have proposed a lexical analysis approach, in which we scan the documents and identify tokens from the abstracts of journal articles. Tokens are considered to be the major source of information in our work. Each journal article can be expressed as a vector of tokens and their weights. The weight of a token is its frequency of occurrence. We use distance as a basis to calculate the contribution of each K neighbor in the class allocation process. And then we calculate the predicted class of the journal article according to the formula given in equation 1 mentioned below. The architecture of the proposed approach is shown in the following Fig. 1.

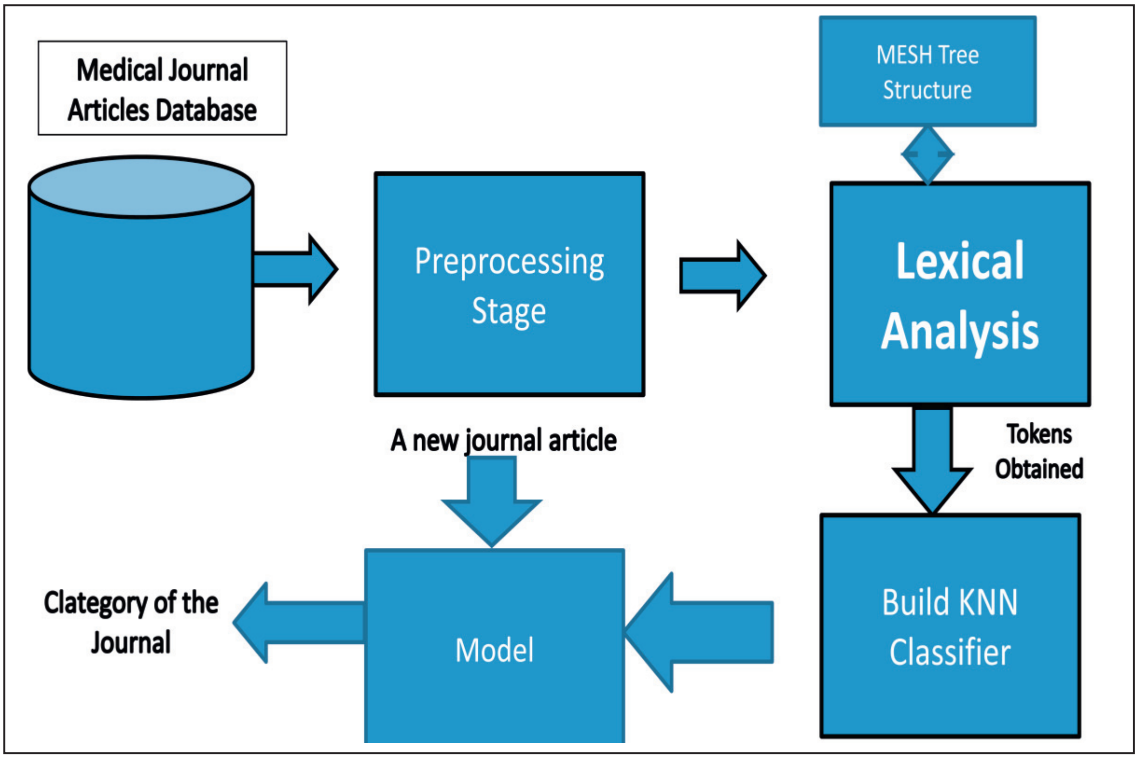

Fig. 1. Architecture of proposed linguistic Approach

The above approach can be explained as follows: In the first part, a collection of journal articles is taken as input. First of all, the articles are scanned and preprocessed. Stop words and special characters like @, : etc. are removed from the abstracts of the articles. The standard list of stop words is given13 .Then the articles are sent to Lexical Analyzer. The job of lexical analyzer is to scan the characters in the given input and group them into tokens. Tokens here are keywords and their related or synonymous words. We have taken a standard list of keywords given as MESH tree Structure[19] . A table is maintained which records the name of the token and its weight. The result of this part is a stream of tokens generated .The frequency of occurrence of tokens is also taken into consideration. It is considered as weight of a token. . Each journal article a is represented as a vector : <w1 (a), w2 (a), w3 (a), w4 (a)...> where wi (a) is the weight of the ith term. That weight is set according to its frequency of occurrence. After this, we have built KNN Classifier using these tokens and their weights. To build KNN Classifier, we use distance as a basis to calculate the contribution of each k neighbor in the class allocation process. The following equation (1) defines the predicted class of a journal article ai belonging to class c as:

$$\mathrm{Pred}_{\mathrm{class}}(\mathrm{c}, \mathrm{j_i}) = \frac{\sum_{\mathrm{k_i} \in \mathrm{K}[\mathrm{Class}(\mathrm{k_i}=\mathrm{c})]} \mathrm{Sim}(\mathrm{k_i}, \mathrm{j_i})}{\sum_{\mathrm{k_i}=\mathrm{K}} \mathrm{Sim}(\mathrm{k_i}, \mathrm{j_i})} \qquad (1)$$

Where Sim is a similarity function which returns a value after comparing an article with its neighbor. That is, we sum up the similarities of each neighbor belonging to a particular class c and divide by all similarities of k neighbors irrespective of the class.

And the last part is testing, in which a new test data arrives, it goes through the whole process and is classified. To compare article j with instance i, we define the Cos Sim function (given in equation (2)) which is defined using our token weight approach as follows:

$$\text{Cos Sim}(i, j) = \frac{S}{\sqrt{A * B}} \tag{2}$$

where S is the number of terms that i and j have in common, A is the number of terms in i and B the number of terms in j.

4. Empirical Evaluation

The empirical evaluation is done on ohsumed test collection[12] compiled by William Hersh. It is a part of the MEDLINE database which is stored by the National Library of Medicine[19]. It contains medical abstracts from the MeSH categories of the year 1991. We have used the dataset[8] in which the first 20,000 documents were divided as 10,000 for training and 10,000 for testing. We have selected the category of 23 cardiovascular diseases. Under this category, the unique abstract number becomes 13,929 (6,286 for training and 7,643 for testing). The purpose of text categorization in this is to allot the documents to one or multiple categories of the total 23 Cardio vascular diseases. A document belongs to a category if it contains at least one indexing term of that category.

The traditional KNN algorithm and LKNN are implemented using JDK 1.6. The input collection of medical journal articles is initially made ready by pre processing. Then tokens are identified by Lexical Analysis module. With the help of tokens and their weights, KNN Classifier is build. A sample original journal article20 from ohsumed collection is shown in Fig. 2a. The article belongs to category 1 with sequence number 988. The pre processed document shown in Fig. 2b. Fig. 2c. shows the internal representation of tokens after lexical analysis.

Possible role of leukotrienes in gastritis associated with Campylobacter pylori. This study was done to evaluate the role of leukotrienes (LTs) in gastritis associated with Campylobacter pylori. Biopsy specimens of gastric mucosa were obtained endoscopically from 18 patients with nonulcer dyspepsia for bacteriological and histological examination and extraction of LTs. There was correlation between the LTB4 level in the mucosa and the degree of gastritis evaluated histologically. The level was higher when infiltration of neutrophils in the gastric mucosa was more extensive. The LTB4 level in mucosa infected with C. pylori was higher than that in noninfected mucosa. These findings suggest that endogenous LTs may be related to the pathogenesis of gastritis associated with C. pylori.

Fig. 2a. Original abstract of a sample journal article

Possible role leukotrienes gastritis associated Campylobacter pylori This study done evaluate role leukotrienes (LTs) gastritis associated Campylobacter pylori Biopsy specimens gastric mucosa obtained endoscopically 18 patients nonulcer dyspepsia bacteriological histological examination extraction LTs correlation LTB4 level mucosa degree gastritis evaluated histologically level higher when infiltration neutrophils gastric mucosa more extensive LTB4 level mucosa infected C. pylori higher noninfected mucosa findings suggest endogenous LTs related pathogenesis gastritis associated C. pylori.

Fig. 2b. Pre processed abstract of the same journal article

Tokens Obtained

- gastric
- Infected
- Associated

Fig. 2c. Tokens obtained from the abstract of the same journal article

5. Performance Evaluation of LKNN over traditional KNN

To evaluate the performance of our proposed algorithm, we have compared it with the traditional KNN algorithm. The most common performance metrics used in text categorization are Recall, Precision and F-measure[22] . They can be calculated as follows: In our experiment, if we define A as the number of true positive samples predicted as positive, B as the number of true positive samples predicted as negative, C as the number of true negative samples predicted as positive and D as the number of true negative samples predicted as negative, then Precision, Recall, F-measure can be expressed as follows.

$$\text{Precision}=A/\,(A+C) \tag{3}$$
$$\text{Recall} = A/(A+B) \tag{4}$$
$$\text{F-measure} = (2^* \text{ Precision} * \text{Recall})/\,(\text{Precision} + \text{recall}) \tag{5}$$

We have done a comparative study of the performance of traditional KNN and LKNN with the different K values. The classification results with the different K values are shown in Fig. 3. We have shown the value of F-measure for traditional KNN and LKNN in Table. 1. The results listed are the best results we get for each algorithm from our experiments. This shows that LKNN classifier performance is much better than traditional KNN classifier in most of the different K values. Another point that can be noted is that as value of K is increasing, LKNN performs better than the traditional KNN. The values are shown in bold in Table. 1. We have taken the values of K from 1 to 20.

Table 1.F-measure values for KNN and LKNN

Values of K	KNN	LKNN
1	1	1
5	0.75	0.77
7	0.76	0.77
10	0.78	0.8
12	0.8	0.8
15	0.83	0.84

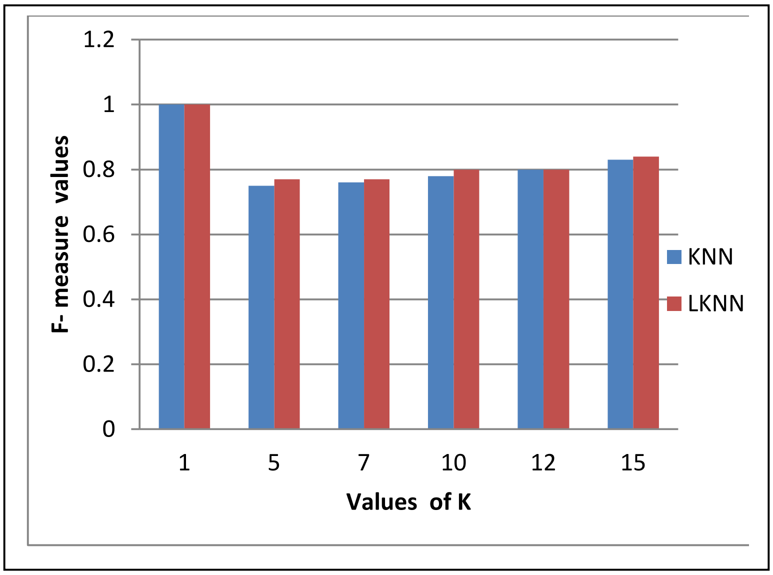

Fig. 3. Comparison between KNN and LKNN over different values of K

6. Conclusion

In this research, we have followed a lexical approach for text categorization in the medical domain. We have proposed an algorithm Lexical KNN (LKNN) which automatically classifies journal articles of medical documents into various categories. The concept of tokens is used to represent a journal article. Each journal article is represented as a vector of tokens. And, further KNN algorithm is used as a text classifier. Our proposed algorithm has outperformed the traditional KNN. This is shown by calculating Recall, Precision and F-measure values. This is a pilot study conducted, in future it can be extended for the complete article or other sections of articles. Also, it can be tested for various other text documents or text datasets.

References

1. Yang Y. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval* 1999: 67–88.
2. Calvo et al. Managing content with automatic document classification. *Journal of Digital Information* 2004; 5:2.
3. Ma L, Shepherd J, Zhang, Y. Enhancing text classification using synopses extraction. *Fourth international conference on web information systems engineering* 2003: 115–124.
4. Soucy P, Mineau GW. A simple k-NN algorithm for text categorization. *First IEEE International Conference on Data Mining* 2001; 28: 647–648.
5. Sasaki M , Kita K. Rule-based text categorization using hierarchical categories. *IEEE international conference on systems, man and Cybernetics* 1998; 3: 2827–2830.
6. Jalam R. , Teytaud O. Kernel-based text categorization. *International joint conference on neural networks* 2001; 3 : 15–19.
7. Schapire RE, Singer Y. Text categorization with the concept of fuzzy set of informative keywords. *IEEE international conference on fuzzy systems*1999; 2 : 609–614.
8. Joachims T. Text categorization with support vector machines: Learning with many relevant features. *Tenth European conference on machine learning* 1998: 137–142.
9. Taeho Jo. Application of table based similarity to classification of biomedical documents. *IEEE International Conference on Granular Computing* 2013: 162 – 166.
10. Mohammed GH A Z , Can A B. ROLEX-SP: Rules of lexical syntactic patterns for free text categorization. *Journal of Knowledge-Based Systems* Elsevier 2011; 24 : 58-65.
11. Zhang Q et al. Machine Learning Methods for Medical Text Categorization. *Pacific-Asia Conference on Circuits, Communications and Systems* 2009: 494 - 497.
12. Hersh W, Buckley C. , Leone T, Hickman D. OHSUMED: An interactive retrieval evaluation and new large text collection for research. 17th *ACM International Conference Research and Development in Information Retrieval* 1994: 192–201.
13. Taneja S, Gupta C, Gureja D and Goyal K. K Nearest-Neighbor Techniques for Data Classification-A Review. *ICCIN 2014;* Delhi, India.
14. Shi L et.al. Text Classification Based on Nonlinear Dimensionality Reduction Techniques and Support Vector Machines. *Third International Conference on Natural Computation* 2007; IEEE Computer Society.
15. Sang Y et. al.An effective discretization method for disposing high-dimensional data. *Journal of Information Sciences* 2014; 270: 73-91.
16 .Shafiei M et.al. Document Representation and Dimension Reduction for Text Clustering. *IEEE 23rd International Conference on Data Engineering Workshop* 2007: 770 - 779 .
17. Jaffali S, Jamoussi S. Principal Component Analysis neural network for textual document categorization and dimension Reduction. *6th International Conference on Sciences of Electronics,Technologies of Information and Telecommunication* 2012: IEEE Xplore.
18. List of Stop words: http://weka.sourceforge.net/doc.dev/weka/core/Stopwords.html.
19 . Mehnert R. Federal agency and federal library reports. 1997.Province, NJ: National Library of Medicine.National Library of Medicine: http://www.nlm.nih.gov/
20. Fukuda T et. al. Possible Role of Leukotrienes in Gastritis Associated with Campylobacter pylori. *Journal of Clinical Gastroenterology*.1990;12:1.
21. Deerwester S et. al. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 1990; 41: 6.
22. Baoli Let. al. An Improved k-Nearest Neighbor Algorithm for Text Categorization. *Proceedings of the 20th International Conference on Computer Processing of Oriental Languages*. Shenyang. China. 2003.