IJCSN International Journal of Computer Science and Network, Volume 3, Issue 4, August 2014
ISSN (Online) : 2277-5420 **www.IJCSN.org**
Impact Factor: 0.274

Medical Document Classification from OHSUMED Dataset

[1] Hira Lal Gope, [2] Pranajit Kumar Das, [3] Dr. Mohammed Jahirul Islam, [4] Md. Hanif Seddiqui

[1] Lecturer, Dept. of CSE, Sylhet Agricultural University, Sylhet-3100, Bangladesh

[2] Lecturer, Dept. of CSE, Sylhet Agricultural University, Sylhet-3100, Bangladesh

[3] Associate Professor, Dept. of CSE, Shahjalal University of Science and Technology Sylhet-3100, Bangladesh

[4] Associate Professor, Dept. of CSE, University of Chittagong, Chittagong-4331, Bangladesh

Abstract - Investigation of biological databases is not straight forward and generically needs identification of Medical Document Classification (MDC) based on hierarchy and relationship between different strata (ontogeny). Thus, MDC remains as a challenging effort. Popularly, earlier text classification has applied flat classifier. However, our research aims to show the text classification in which we opt to assess the hierarchical organization of classes or categories. In order to fulfill the aim of our research, we are considering the human disease hierarchical structure of human disease ontology with the help of simple relation from biomedical text abstracts and the ontology learning. We conducted experiments to evaluate the effects of different representations by measuring the change in classification performance with MEDLINE documents from the OHSUMED dataset. This research suggest a hierarchical classification method employing the hierarchical concept structure for classifying biomedical text abstracts by using Hidden Markov Model method (HMM). Present study demonstrates how a large number of biomedical articles are divided into quite a few subgroups in a hierarchy describing ontogeny.

Keywords - *Biomedical articles, HMM, Hierarchical classifier, Human diseases, Medline documents, Ontology learning.*

1. Introduction

Biomedical literature aims to select relevant articles to a specific issue from large data volume for document classification system. However, classifying biomedical literature [1, 2] becomes one of the challenging tasks when the number of categories grows to a significantly large number. This is due to the fact that, it will become much more difficult to browse and search the categories. One way to solve this problem is to organize the categories into a hierarchy. Text classification [2] system on biomedical literature aims to select relevant articles to a specific issue from large data volume. Hierarchical structures identify the relationships of dependence between the categories and provide a valuable information source for many problems. We are confident that, by introducing a hierarchy to a huge collection of biomedical text abstracts, it can help us to classify these abstracts according to their specific category. Recently, several researchers have investigated the use of hierarchies for text classification [1], yet, only little attention has been done to apply to the biomedical literature. For that reason, in this research, we are exploring the application of hierarchical structure for classifying a collection of biomedical text abstracts that related to human diseases. However, we have no systematic method to build a hierarchical classification system that performs well with large collections of practical data.

To overcome this problem, we propose a framework for hierarchical classification method with the help of ontology and utilizing the techniques of ontology alignment. In this research, our aim is to investigate the method for constructing ontology learning and human disease ontology for ontology alignment and hierarchical. In [3], Li, et.al said, hierarchical structure identify the relationships of dependence between the categories and provide a valuable information source for many problems. In order to achieve the research goal, we will conduct the experiments using the OHSUMED dataset and a subset of biomedical text abstracts from MEDLINE database that related to human diseases. Theoretically, our approach is capable to classify more relevant concepts or categories for a collection of biomedical text abstracts by applying ontology alignment. In [4], Singh and Nakata stated that, flat classification approach is suitable when a small number of categories are defined. However, in areas such as search result classification, where the retrieved documents can belong to several different categories, flat classification becomes inefficient and hierarchical classification is preferred.

2. Background

Hierarchical classification refers to the assignment of one or more suitable categories from a hierarchical category space to a document. Now-a-days [1], so many researchers have found out the use of hierarchies for text classification. But still now, a bit attention has been done to apply to the biomedical literature. That's why, this research is motivated for exploring the application of hierarchical structure for classifying a collection of biomedical text abstracts that related to human diseases. However, we are in the depth of limitations of a proper systematic method for building a hierarchical classification system that performs at large with huge collections of practical data. To find out a solution of this problem, by utilizing the techniques of ontology alignment, we suggest a framework for hierarchical classification method.

3. State-of-the-Art

To understand the underlying mechanisms of human diseases is one the challenging limitations in biomedical research. One of the main efforts of the research in automated text classification area has been to adopt and enhance machine learning algorithms, such as decision trees, nearest neighbor, naive-bayes, neural networks, and Support Vector Machines (SVM) [5]. Another focus has been to compare the performance of existing classifiers [6];[7]. From this body of research, the high dimensional nature of text data has been shown to be the main reason for the bad performance of many classifier methods. After investigating the classifier options listed above, we decided to use SVM in our experiments because its performance has been superior for high dimensional feature representations, such as those necessary for text.

4. Related Work

Generally, text classification can be considered as a flat classification technique, where the documents are classified into predefined categories and there is no relationship specified between the categories. However, in areas such as search result classification, where the retrieved documents can belong to several different categories, at classification becomes inefficient and hierarchical classification is preferred. Contrary to at classification, hierarchical classification can be defined as a process of classifying documents into a hierarchical organization [8] of classes or categories. In hierarchical structure, we can identify and provide the relationships of dependence between the classes or categories. A few hierarchical classifications methods have been proposed recently. In most of the hierarchical classification methods, the categories are organized in tree like structures. In addition, hierarchical classification and ontology also has attracted the attention of researchers. Therefore, in this research, we explore the application of hierarchical structure and we propose the use of ontology, especially ontology alignment for classifying of biomedical text abstracts. Eventually, by utilizing the techniques of ontology alignment in our approach, we can produce more relevant concepts for biomedical hierarchical classification.

5. Proposed System

Here we show our proposed system in **Fig. 1,**

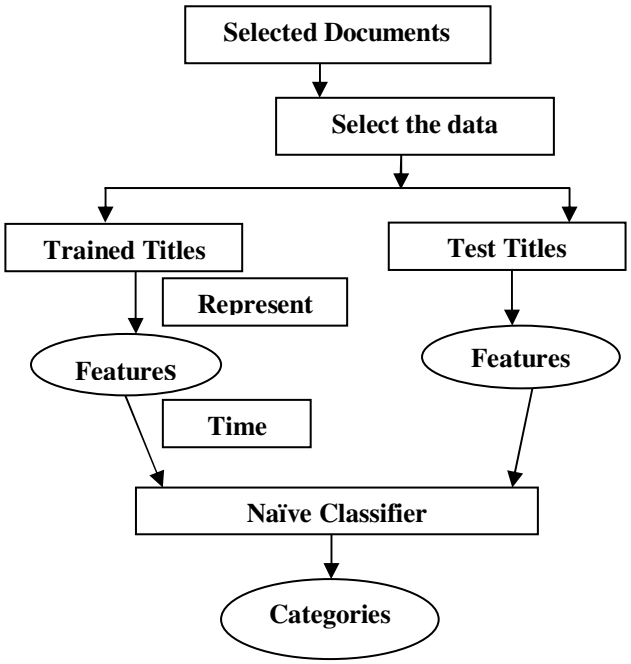

Fig. 1. Naïve Bayes' Classification outline

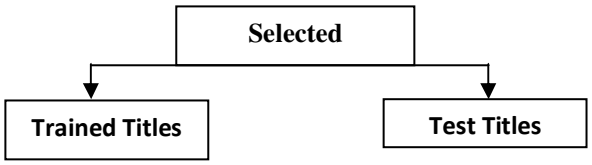

Fig. 2. Basic parts of an ontological system

Ontological System first divides the selected documents title in two basic parts (**see Fig. 2**). Then both Test and Training will classified into features which is the output in Text format. After being classified into text then ontological knowledge base is applied to train the Text format dataset. Finally these dataset will fit to classifier to classify the dataset.

6. Methodology

The main aspects of this research are to develop MDC with good accuracy. The research is ongoing, and some proposals are under consideration as complements to the currently planned approach. Using Hidden Markov Model we have designed following algorithmic steps:

- Completing list of tags that was necessary for run our program correctly.
- Part-of-speech tagging (POS tagger) that was help to identified noun, verb, and adjective and so on.
- Word chunking that was helped us to identify noun and verb phrases.
- Word and chunk frequency which help us to find out any phrase how many time occurs.
- Document similarities which we measure it from word frequency and chunk frequency.
- Document classifications are one of important steps in document mining.

6.1 Hidden Markov Model

A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. An HMM can be considered as the simplest dynamic Bayesian network. In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence

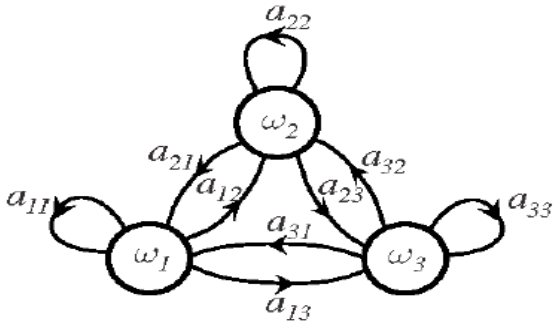

Fig. 3. : Hidden Markov Chain

of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; even if the model parameters are known exactly, the model is still 'hidden'. Hidden Markov models are especially known for their application in temporal pattern recognition [9] such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, partial discharges and bioinformatics. A hidden Markov model can be considered a generalization of a mixture model where the hidden variables (or latent variables), which control the mixture component to be selected for each observation, are related through a [11, 12, 13] Markov process rather than independent of each other. The working procedure of HMM is given here as a mathematical scratch (**See Fig. 3**).

6.2 Part-of-Speech Tagging

In corpus linguistics, part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category[7] disambiguation, is the process of marking up the words in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, and so on. Part-of-speech tagging is a process whereby tokens are sequentially labeled with syntactic labels, such as "finite verb" or "gerund" or "subordinating con-junction". This tutorial shows how to train a part-of-speech tagger and compile its model to a file, how to load a compiled model from a file and perform part- of-speech tagging, and finally, how to evaluate and tune models.

6.3 Phrase Chunking

Phrase chunking is a natural language process [10] that separates and segments a sentence into its sub constituents, such as noun, verb, and prepositional phrases. It is the process of recovering the phrases (typically base noun phrases and verb phrases) constructed by the part-of-speech tags. For instance, in the sentence John Smith will eat the beans. There is a proper noun phrase John Smith, a verb phrase will eat and a common noun phrases the beans. Note that this notion of phrase may not line up with any theoretically motivated linguistic analysis.

6.4 Document Similarities

Document Similarities become one of the important and challenging tasks [11].We measure it from word frequency and chunk frequency. For now, the system computes only a single numeric value for each pair of documents in a given set. The most common use case is to discover which documents are similar to the given document. This value represents the number of chunks which these two documents have in common (not taking the possible hash function collisions into the account). The most common

use case is to discover which documents are similar to the given document (e. g. a newly imported thesis).

6.5 Classifying Documents

After representing the titles we trained our classification method with the training titles and tested its performance with the test titles. We picked Hidden Markov Model (HMM) as our classification method due to its superior performance with text compared to their methods. Then it trains the classifiers with the training sets, classifies each test set with the corresponding trained classifier.

6.6 Dataset

OHSUMED means An Interactive Retrieval Evaluation and New Large Text Collection for Research. The OHSUMED dataset is a subset of clinical paper abstracts from the Medline database consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987-1991). It consists of the 23 Medical Subject Headings (MeSH) diseases categories. While, for testing ontology (we named it as human disease ontology), we randomly downloaded 100 biomedical text abstracts related to human diseases that available in the Medline database. Afterward, we extract the features (noun phrases and verb phrases) by performing part-of-speech (POS) tagging and phrase chunking. POS tagging is a task of assigning POS categories to terms from a predefined set of categories. For this purpose, we employed Hidden Markov Model (HMM). Taggers based on the HMM technology currently appear to be in the lead. Meanwhile, phrase chunking is the process of recovering the phrases (typically base noun phrases and verb phrases) constructed by the part-of-speech tags. Finally, we employed the relevant noun phrases (as concepts) and verb phrases (as relations) for constructing human disease ontology and ontology learning.

7. Conclusion

Here we are working on ontology-based hierarchical classification to evaluate the performance of biomedical text abstract classification. One of the reasons for this performance increase was that we used an NLP tool that was designed purely for medical phrase identification and a medical knowledge-base in our bag-of-phrases representation.

8. Future Directions

Our future target is to propose more efficient approaches for identifying related concepts and discovering more complex relations between known concepts for refining and enriching our ontology learning and human diseases ontology in order to improve the performance of hierarchical text classification. In future we should applied Support Vector Machine (SVM) to find out better results. Classification in SVM is an example of Supervised Learning. Known labels help indicate whether the system is performing in a right way or not. This information points to a desired response, validating the accuracy of the system, or be used to help the system learn to act correctly. A step in SVM classification involves identification as which are intimately connected to the known classes. This is called feature selection or feature extraction.

Acknowledgements

In this work I am grateful to Md. Hanif Seddiqui, PhD. (Japan), Associate Professor Department of Computer Science & Engineering, University of Chittagong, Bangladesh for his idea. I have had inspired by him as well as the necessity of current world.

References

[1] Binti Dollah, Rozilawati, Md. Hanif Seddiqui, and Masaki Aono. "The effect of using hierarchical structure for classifying biomedical text abstracts."

[2] F. M. Couto, B. Martins and M. J. Silva, Classifying biological articles using web sources, Proceedings of the 2004 ACM symposium on Applied Computing, pp. 111-115 (2004).

[3] T. Li, S. Zhu and M. Ogihara, Hierarchical document classification using automatically generated hierarchy, Journal of Intelligent Information Systems, Springer Netherlands, Vol.29, No.2, pp. 211-230 (2007).

[4] Singh, Amrish, and Keiichi Nakata. "Hierarchical classification of web search results using personalized ontologies." Proceedings of the 3rd International Conference on Universal Access in Human-Computer Interaction, HCI International. Vol. 2005. 2005.

[5] SEBASTIANI, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34, 1-47.

[6] Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998, November). Inductive learning algorithms and representations for text categorization. In Proceedings of the seventh international conference on Information and knowledge management (pp. 148-155). ACM.

[7] Lewis, D. D., Schapire, R. E., Callan, J. P., & Papka, R. (1996, August). Training algorithms for linear text classifiers. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 298-306). ACM.

[8] Couto, Francisco M., Bruno Martins, and Mário J. Silva. "Classifying biological articles using web resources." Proceedings of the 2004 ACM symposium on Applied computing. ACM, 2004.

[9] Lewis, David D. (1992, June). An evaluation of phrasal and clustered representations on a text categorization task. In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 37-50). ACM.

[10] Mao, Wenlei, and Wesley W. Chu. "Free-text medical document retrieval via phrase-based vector space model." Proceedings of the AMIA Symposium. American Medical Informatics Association, 2002.

[11] Wilcox, Adam, George Hripcsak, and Carol Friedman. "Using knowledge sources to improve classification of medical text reports." *KDD-2000*. 2000.

[12] UmutTosun HIDDEN MARKOV MODELS TO ANALYZE USER BEHAVIOUR IN NETWORK TRAFFIC Bilkent University 06800 Bilkent, Ankara, Turkey.

[13] Biological Data Working Group Federal Geographic Data Committee and USGS Biological Resources Division.

Biography

[1] I was born and grew up in Bangladesh. I am a recent graduate of the department of Computer Science and Engineering at the University of Chittagong, Bangladesh. I joined as a Lecturer in Computer Science and Engineering department at Sylhet Agricultural University, Bangladesh. I am a believer in life-long learning and I am passionate about the natural language processing, semantic knowledge base, compiler design, operating system and bioinformatics.