2012 International Conference on Information and Computer Networks (ICICN 2012)
IPCSIT vol. 27 (2012) © (2012) IACSIT Press, Singapore

Ontology for Medical Document Classification

Sarwar kamal [1], Sonia Nimmy [2]

[1] Lecturer, Computer science and engineering BGC Trust University Bangladesh Chittagong.
sarwar.saubdcoxbazar@gmail.com

[2] Lecturer, Computer science and engineering BGC Trust University Bangladesh Chittagong.
nimmy_cu@yahoo.com

Abstract. In the age of information superhighway the volume of documents in Bioinformatics is growing dramatically and every day it is becoming difficult to organize such size of data. We have to remember that Bioinformatics is Information based and data intensive science. So it is very important for us that we have to keep proper track and design technique according to the increase of data in our everyday life throughout the world. New data management and managing process are essential in the field of medical information science. The identification of Medical Document Classification has become challenging due to its division of sub groups in a hierarchy. In the past text classification has applied at classifier. However, in this paper we will show the text classification in which we will assess the hierarchical organization of classes or categories. In order to finish this work we will consider the human disease hierarchical structure of human disease ontology with the help of simple relation from biomedical text abstracts and the ontology learning. Medical document classification (MDC) has wide-spread problem with many applications and challenging tasks. The aim of this study is to investigate how a large number of biomedical articles are divided into quite a few subgroups in a hierarchy. MDC is the process of transforming descriptions of medical diagnoses and procedures into universal medical code numbers. Here we propose a hierarchical classification method where we employ the hierarchical concept structure for classifying biomedical text abstracts by using Hidden Markov Model.

Keywords: Ontology, MDC, HMM, Hierarchical, Biomedical Text, Superhighway, Bioinformatics.

1. Introduction

Document classification system on biomedical literature aims to select relevant articles to a specific issue from large data volume. 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. 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 order to achieve the research goal, we will conduct the experiments using the EVOLUTIONARY BIOLOGICAL dataset and a subset of biomedical text abstracts from MEDLINE database that related to human diseases.

2. Bioinformatics and Ontology

Biologists need knowledge to perform their work, often using a pre-existing item of knowledge to make inferences about the item under investigation. The most common example of this within molecular biology is

the use of sequence comparison to infer the function of a novel protein sequence. The reasoning is that if a sequence of unknown function is highly similar to a sequence of known function, then it is probable that the novel sequence also has that function. So, rather than using a rule, law or equation to find the function of a protein, a biologist uses the knowledge that a similar sequence has a known function to make a judgment about the function of the new sequence. This is why it is sometimes said that biology is a knowledge-based, rather than an axiom-based discipline. Modern biologists also need knowledge for communication. Biology is a data-rich discipline, which is available as a fund of knowledge by which biologists generate further knowledge. This knowledge is stored in thousands of databases, many of which need to be used in concert during an investigation. Knowledge is vital in two respects during this process. For instance, when using more than one data store or analysis tool, a biologist needs to be sure that knowledge within one resource can be reliably compared with another. A prime example is the differing uses of the term gene within the community. In one database, gene may be defined as the coding region of DNA; in another as DNA fragment that can be transcribed and translated into a protein and DNA region of biological interest with a name and that carries a genetic trait or phenotype in a third. Being able to conform to a common definition or reason about the differences between definitions, in order to reconcile databases, would be advantageous. The second need for knowledge is to define and constrain data within a resource. Biological data can be very complex; not only in the type of data stored, but in the richness and constraints working upon relationships between those data. When designing a database it is useful to be able to describe what values can be specified for which attributes under which conditions. This is the encapsulation of biological knowledge within database schema. It is impossible for one biologist to deal with all the knowledge within even one sub domain of their discipline. The arrival of whole genomes and the knowledge they contain only exacerbates the situation. There is, therefore, a need for systems that can apply the domain experts' knowledge to biological data. It is not envisaged that such systems could ever perform better than human experts; however, they could play a crucial role in helping process data to the point where human experts could again apply their knowledge sensibly. This raises numerous questions, in particular regarding how knowledge can be captured to make it available and useful within computer applications. Knowledge can be captured and made available to both machines and humans by ontology. The premise for the need for ontologies within bioinformatics is the need to make knowledge available to that community and its applications.

3. 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[2] 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.

4. Our Proposed System

According to our goal our Ontological System first divides the selected documents title in two basic parts:

1. Test Title
2. Training titles

Then both Test and training will classified into features which is the output in Text format. After being classified into text then ontological knowledgebase is applied to train the Text format dataset.

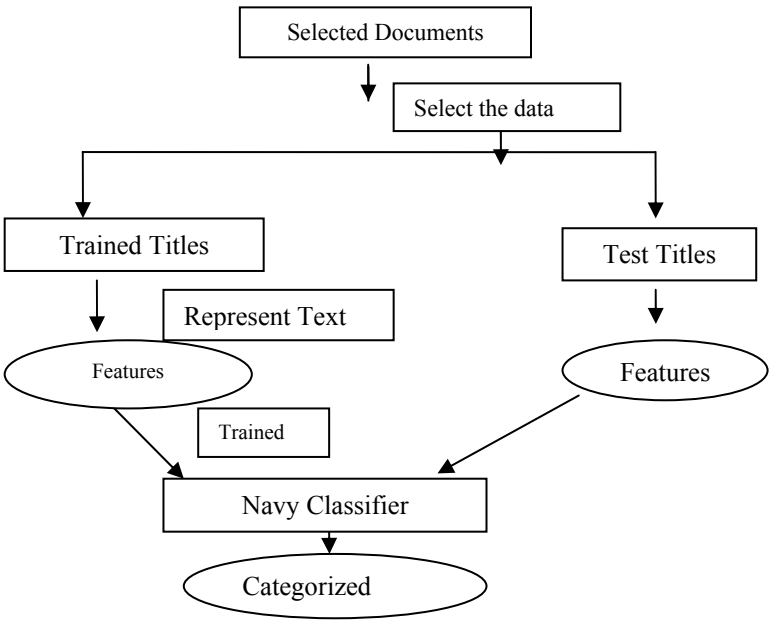


Fig 1: The Naive Bayes' classification outline

Finally these dataset will fit to classifier to classify the dataset. Here we have chosen Naive Bayes'. Here is an illustration of our results of Evolutionary Biological Dataset classification Within the evolutionary biological dataset we have to select first documents title which is very important and mandatory I for Naive Bayes'Statistical classification.

5. 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:

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

5.1. 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 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 [5] 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 [10, 11, 12] Markov process rather than independent of each other. The working procedure of HMM is given here as a mathematical scratch.

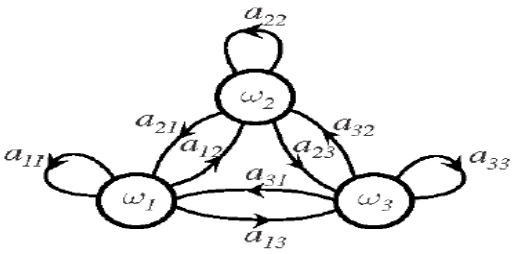


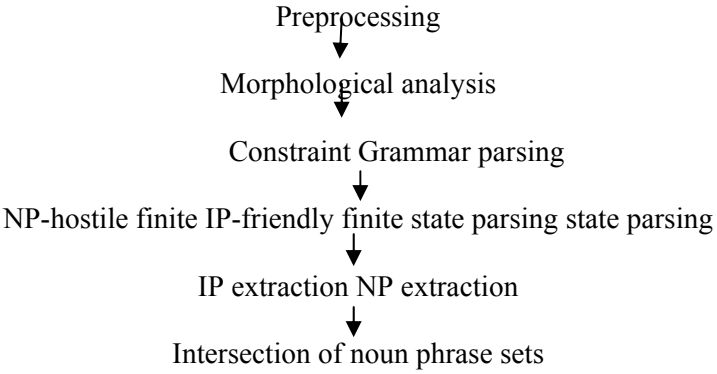
Fig 2: Hidden Markov Chain

5.2. Part-of-Speech Tagging

In corpus linguistics, part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category[6]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.

5.3. Phrase Chunking

Phrase chunking is a natural language process [7] 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 phrase, the beans. Note that this notion of phrase may not line up with any theoretically motivated linguistic analysis.



5.4. Document Similarities

Document Similarities become one of the important and challenging tasks[11] . We measure it from word frequency and chunk frequency. This inverted index is then used for finding similarities in a given document base. For now, the system computes only a single numeric value for each pair of documents in a given set. 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). We postpone the computation of the actual similar passages of the text to the time when the user wants to see them.

5.5. Classifying Documents

After representing the titles with bag-of-words, bag-of-medical-phrases, and hybrid approaches, 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 other methods. HMMs are based on the Structural Risk[9] Minimization principle from computational learning theory. They are linear classifiers that try to find a hyper-plane that maximizes the margin between the hyper-plane and the given positive and negative examples. For our text classification case, a medical document can be assigned to more than one Mesh category; thus, this problem can be viewed as a series of binary classification problems, one for each category, rather than as a multi-class classification problem. In our system, we used the latest version of SVM-Light, a very commonly used implementation of SVM developed by Thorsten Joachims [8]. Our system generates a training and test set for each category by labeling all the training or test titles in the category as positive examples and the rest of training or test titles as negative examples. Then it trains the classifiers with the training sets, classifies each test set with the corresponding trained classifier, and measures the overall performance by calculating precision, recall, and F1-Score metrics for the classification results. Here is the depiction of biological dataset classification from biological dataset [12].

Table 1:Depicts the classification of documents and text representation

Classification of documents		Text representation
Title and dataset Classification = Citation + Elaboration + Time Period + Nature of the Title + **0{**Selected Documents**} 1** + Titles + 0{Headings}1 + Sub heading + User Manual + (Point of NOTES) + (1{Graphical View}n) + (Data Types) + (Security) +	(Native Data Set Environment) + (1{Cross Reference}n) **+** 0{Analysis}n Citation = 1{Origin}n + Date + (Title Time) + Title + 0{Edition}1 + Biological Dataset Form + 0{Details Information}1 + 0{Publication Information}1 + 0{Other Citation if any}1 + (1{Online Linker}n) +	CGM" Computer Graphics Metafile "EPS" Encapsulated Postscript format "GIF" Graphic Interchange Format "JPEG" .Joint Photographic Experts Group format "PBM" Portable Bit Map format "PS" Postscript format "TIFF" Tagged Image File Format "XWD" X-Windows Dump Extended Domain: If a definition of the type is not contained in the table below, use a standard mime type extension if possible, before creating a new name. “AIF” Audio Interchange File Format “ASF” Advanced Streaming Format “AU” Sun audio format “AVI” Audio Video Interleave format “MID” Musical Digital Interface format MOVIE” SGI movie video format “MP3" MP3 music format “MPEG” Moving Picture Experts Group video format “MPGA” MPEG audio format “PNG” Portable Network Graphics format “PPT” PowerPoint presentation

6. Important Information

In this paper, we presented the effects of using different representation approaches on the overall performance of medical document 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. We have reported the results from our classification experiments with the BIOLOGICAL data-set. With absolute rejecter classifiers, we have shown that the BIOLOGICAL dataset does not include enough training data for many categories, especially for the more general and broad ones. We plan to continue to our work with larger and more representative datasets gathered from EVOLUTIONARY BIOLOGICAL DATASET.

7. Acknowledgements

In this work I m grateful to Dr.Hanif Seddiqui Associate Professor Laboratory 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.

8. References

[1]. Binti Dollah, R., Seddiqui, M. & Aono, M.. The effect of using hierarchical structure for classifying biomedical text abstracts.

[2].Couto, F., Martins, B. & Silva, M. (2004). Classifying biological articles using web resources. In Proceedings of the 2004 ACM symposium on applied computing.

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

[4]Joachims, T. (2002). Learning to classify text using support vector machines: Methods, theory, and algorithms. Computational Linguistics, 29, 656{664.

[5]. Lewis, D. (1992). 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 , 37{ 50, ACM

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

[7]. Mao, W. & Chu, W. (2002). Free-text medical document retrieval via phrasebased vector space model. In Proceedings of the AMIA Symposium, 489, American Medical Informatics Association.

[8]. Ruiz, M. & Srinivasan, P. (1999). Hierarchical neural networks for text categorization (poster abstract). In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 281{282, ACM.

[9]. Sebastian, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34

[10]. Umut Tosun HIDDEN MARKOV MODELS TO ANALYZE USER BEHAVIOUR IN NETWORK TRAFFIC Bilkent University 06800 Bilkent, Ankara, Turkey

[11]. Wilcox, A., Hripcsak, G. & Friedman, C. (2000). Using knowledge sources to improve classification of medical text reports. In KDD-2000, Citeseer.

[12]. Biological Data Working GroupFederal Geographic Data Committee and USGS Biological Resources Division.