3A3-02

The effect of using hierarchical structure for classifying biomedical text abstracts

Rozilawati binti Dollah Md Hanif Seddiqui Masaki Aono

Department of Computer Science and Engineering
Toyohashi University of Technology

Classifying biomedical literature becomes one of the important and challenging tasks lately, due to the fact that a large number of biomedical articles are divided into quite a few subgroups in a hierarchy. It causes difficulties for the researcher to effectively and efficiently organize and retrieve relevant information from the database. In the past, most approaches used in text classification task have applied flat classifiers that ignore the hierarchical structure and treat each concept separately. Therefore, in this paper, we present an exploration of the application of hierarchical structure for classifying a collection of biomedical text abstracts downloaded from the Medline database. To accomplish this goal, we construct the human disease hierarchical structure or human disease ontology with some simple relations from biomedical text abstracts and the ontology learning. Subsequently, we enrich the ontology learning before adapting it to ontology alignment process for classification purpose. Eventually, our approach can identify more relevant concepts for classifying biomedical text abstracts by utilizing ontology especially, the techniques of ontology alignment.

1. Introduction

Text classification system on biomedical literature aims to select relevant articles to a specific issue from large corpora [1]. However, classifying biomedical literature 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. In [2], Li, et.al said, hierarchical structure 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, 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 classification purpose. 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.

The rest of the paper is organized as follows. In Section 2, we give an overview of hierarchical text classification. The framework of hierarchical classification method is discussed in Section 3. Finally, Section 4 concludes with a summary and suggestions for future work.

2 Hierarchical Text Classification

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. In [3], 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.

Contrary to flat classification, hierarchical classification can be defined as a process of classifying documents into a hierarchical organization of classes or categories. In hierarchical structure, we can identify and provide the relationships of dependence between the classes or categories.

A few hierarchical classification 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.

3 The Framework of Hierarchical Classification Method

This section describes the dataset and the method employed in our research. Lately, various classification methods are proposed for classifying biomedical literature. However, in our research, we explore the use of hierarchical structure or ontology for biomedical text classification. The features or concepts in ontology can be used to index the biomedical text abstracts for improving the accuracy of classification performance and also the result of searching relevant documents.

Generally, ontology can be defined as a hierarchical organization of concepts where the textual documents assigned to each concept explain its semantics and the directed hierarchical structure provides an understanding of the relationships between them [3]. In our research, we implemented Anchor-Flood algorithm (AFA) [4] for aligning human disease ontology and ontology learning. During ontology alignment, AFA will collect small blocks of concepts and related relations for classification purpose. Figure 1 below shows the proposed framework that implemented in our research.

3.1 Dataset

We use the OHSUMED dataset for constructing and enriching the ontology learning. The OHSUMED dataset is a subset of

Author : Rozilawati binti Dollah, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi, 0532-44-6764, rozeela@kde.cs.tut.ac.jp

clinical paper abstracts from the Medline database, from year 1987 through 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.

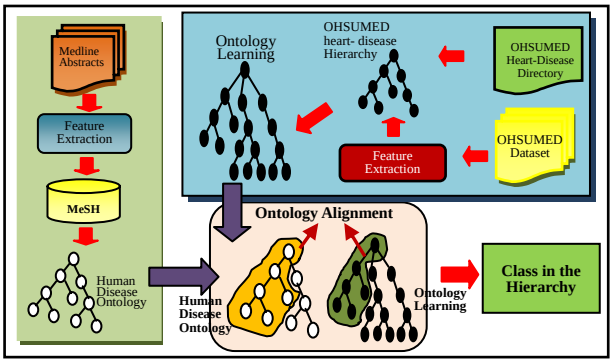

Fig.1: A framework for biomedical hierarchical classification

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.

3.2 Human disease ontology

In our research, we create human disease ontology with some simple relations based on the selected noun phrases and verb phrases that extracted from biomedical text abstract. For this purpose, we refer to Medical Subject Headings (MeSH) tree structures for identifying heading and subheading of hierarchical grouping for selected noun phrases. While, the discovered verb phrases between noun phrases are used to build a network of concepts and relations for human disease ontology.

3.3 Ontology learning

Ontology learning techniques can be divided in constructing ontology manually and extending existent ontology. In our research, we construct ontology based on ontology learning of OHSUMED heart-disease hierarchy by referring to the OHSUMED heart-disease directory. Subsequently, we enrich the concepts in OHSUMED heart-disease hierarchy by inserting new concepts and relations (based on selected noun phrases and verb phrases that extracted from OHSUMED dataset) into the OHSUMED heart-disease hierarchy.

The important task in enriching the OHSUMED heart-disease hierarchy is to derive meaningful and important concepts and relations from OHSUMED dataset. In most researches have been conducted, many researchers were employing only noun phrases as concepts for ontology building and ignored any ontological relations between concepts. Therefore, for ontology construction, we extract noun phrases and verb phrases from OHSUMED dataset through POS tagging and phrase chunking process.

3.4 Ontology Alignment

The purpose of ontology alignment in our research is to match between two ontologies for collecting neighboring concepts and relations to classify biomedical text abstracts. In this research, we perform ontology alignment using Anchor-Flood algorithm (AFA) to align concepts and relations of the human disease ontology and "enriched" OHSUMED heart-disease hierarchy. During ontology alignment process, AFA will look for the similarity among the neighbors based on terminological alignment and structural alignment.

3.5 Ontology-based Hierarchical Classification

In our research, we exploit the hierarchical structure and their relations together with similarity measures in order to identify and predict more specific concept or category for biomedical text classification. Therefore, starting from the hierarchical concept leaves, our approach repeatedly traverses towards the neighboring concepts, including children, siblings and parents, until we find the concept that has the largest similarity and relevance in the hierarchical structure. Eventually, we evaluate the more specific concepts based on the similarity between the classifiable abstracts and categories.

4 Discussion and Conclusion

In this paper, we described an approach that utilized the techniques of ontology alignment for improving the performance of biomedical text classification. We investigated and explored the effect of the application of hierarchical structure for classifying a collection of biomedical text abstracts that downloaded from the Medline database. In order to achieve this aim, firstly, we constructed OHSUMED heart-disease hierarchy and human disease ontology. Afterward, we enriched OHSUMED heart-disease hierarchy before adapting it to the ontology alignment process for classification purpose. During ontology alignment process, we collected all the aligned concepts together with its neighbor concepts and relations to classify biomedical text abstracts. We used hierarchical structure and their relations in human disease ontology to predict more specific category for biomedical text classification.

We are working on ontology-based hierarchical classification to evaluate the performance of biomedical text abstract classification. We are strongly confident that, our approach can improve the performance of biomedical text classification significantly, then produce more relevant concepts for classifying biomedical text abstracts with the help of ontology and utilizing the techniques of ontology alignment.

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.

Acknowledgements

This study was partially supported by Global COE Program "Frontiers of Intelligent Sensing" from Japan's Ministry of Education, Culture, Sports, Science and Technology.

References

[1] 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).

[2] 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).

[3] A. Singh and K. Nakata, Hierarchical classification of web search results using personalized ontologies, In Proceedings of the 3rd International Conference on Universal Access in Human-Computer Interaction, HCI International 2005 (2005).

[4] M. Seddiqui and M. Aono, An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size, Journal of Web Semantics: Science, Services & Agents on the World Wide Web, pp.344-356 (2009).