Ontology based Approach for Classifying Biomedical Text Abstracts

Rozilawati Binti Dollah rozeela@kde.cs.tut.ac.jp
Dept. of Electronic and Information Engineering
Toyohashi University of Technology
Hibarigaoka, Tempaku-cho,
Toyohashi-shi, Aichi, 441-8580 Japan

Masaki Aono aono@kde.cs.tut.ac.jp
Dept. of Computer Science and Engineering
Toyohashi University of Technology
Hibarigaoka, Tempaku-cho,
Toyohashi-shi, Aichi, 441-8580 Japan

Abstract

Classifying biomedical literature is a difficult and challenging task, especially when a large number of biomedical articles should be organized into a hierarchical structure. Due to this problem, various classification methods were proposed by many researchers for classifying biomedical literature in order to help users finding relevant articles on the web. In this paper, we propose a new approach to classify a collection of biomedical text abstracts by using ontology alignment algorithm that we have developed. To accomplish our goal, we construct the OHSUMED disease hierarchy as the initial training hierarchy and the Medline abstract disease hierarchy as our testing hierarchy. For enriching our training hierarchy, we use the relevant features that extracted from selected categories in the OHSUMED dataset as feature vectors. These feature vectors then are mapped to each node or concept in the OHSUMED disease hierarchy according to their specific category. Afterward, we align and match the concepts in both hierarchies using our ontology alignment algorithm for finding probable concepts or categories. Subsequently, we compute the cosine similarity score between the feature vectors of probable concepts, in the "enriched" OHSUMED disease hierarchy and the Medline abstract disease hierarchy. Finally, we predict a category to the new Medline abstracts based on the highest cosine similarity score. The results obtained from the experiments demonstrate that our proposed approach for hierarchical classification performs slightly better than the multi-class flat classification.

Keywords: Biomedical Literature, Feature Selection, Hierarchical Text Classification, Ontology Alignment

Corresponding author. Address: Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, UTM Skudai, 81310 Johor Bahru, Johor, Malaysia.

1. INTRODUCTION

Text classification is the process of using automated techniques to assign text samples into one or more set of predefined classes [1], [2]. Nonetheless, text classification system on biomedical literature aims to select relevant articles to a specific issue from large corpora [3]. Recently, classifying biomedical literature becomes one of the best challenging tasks due to the fact that a large number of biomedical articles are divided into quite a few subgroups in a hierarchy. Many researchers have attempted to find more applicable ways for classifying biomedical literature in order to help users to find relevant articles on the web. However, most approaches used in text classification task have applied flat classifiers that ignore the hierarchical structure and treat each concept separately. In flat classification, the classifier assigns a new documents to a category based on training examples of predefined documents.

Generally, text classification can be considered as a flat classification technique, where the documents are classified into a predefined set of flat categories and no relationship specified between the categories. Singh and Nakata [4] stated that the flat classification approach was suitable when a small number of categories were defined. Nevertheless, due to the increasing number of published biomedical articles on the web, the task of finding the most relevant category for a document becomes much more difficult. Consequently, flat classification turns out to be inefficient, while hierarchical classification is more preferable. Contrary to flat classification, hierarchical classification can be defined as a process of classifying documents into a hierarchical organization of classes or categories based on the relationship between terms or categories. In hierarchical classification, a new document would be assigned to a specific category based on the concepts and relationships within the hierarchy of predefined classes. Many large text databases, such as Yahoo and Google are organized into hierarchical structure, which would help the users searching and finding relevant articles or information easier.

Lately, the use of hierarchies for text classification has been widely investigated and applied by many researchers. For example, Pulijala and Gauch [5] and Gauch et al. [6] have reported that they classified the documents during indexing which can be retrieved by using a combination of keyword and conceptual match. Ruiz and Srinivasan [7] have proposed a text categorization method based on the Hierarchical Mixture of Expert (HME) model using neural networks. Li et al. [8] have proposed another approach of hierarchical document classification using linear discriminant projection to generate topic hierarchies. In addition, Deschacht and Moens [9] have proposed an automatic hierarchical entity classifier for tagging noun phrases in a text with their WordNet synset using conditional random fields (CRF). Meanwhile, Xue et al. [10] have developed a deep-classification algorithm to classify web documents into categories in large-scale text hierarchy.

Additionally, several statistical classification methods and machine learning techniques have been applied to text and web pages classification including techniques based on decision tree, neural network [7] and support vector machine (SVM) [2], [11], [12], [13]. SVM has been prominently and widely used for different classification task, in particular for document classification. For instance, in [2], Sun and Lim have developed a top-down level-based hierarchical classification method for category tree using SVM classifiers. Afterward, they have evaluated the performance of their hierarchical classification method by calculating the category-similarity measures and distance-based measures. Nenadic et al. [11] have used SVM for classifying the gene names from the molecular biology literature. Meanwhile, Wang and Gong [9] have used SVM to distinguish between any two sub-categories under the same concept or category for web page hierarchical classification. They have used the voting score from all category-to-category classifier for assigning a web document to a sub-category. And they have reported that their method can improve the performance of imbalanced data. Dumais and Chen [13] have reported that they used the hierarchical structures for classifying web content. In their research, they have employed SVM to train second-level category models using different contrast sets. Then, they have classified a web content based on the scores that were combined from the top-level and second-level model. In [14], Liu et al. have analyzed the scalability and the effectiveness of SVM for classifying a very large-scale taxonomy using distributed classifier.

Although many classification methods have been proposed in various domains in recent years such as in [5], [6], [7], [8], [9], [10], [11], [12], [13], [14] and [15], only little attention has been paid to the hierarchical classification of biomedical literature. Towards this effort, in this paper, we propose a hierarchical classification method where we employ the hierarchical ‘concept’ structure for classifying biomedical text abstracts by using ontology alignment algorithm that we have developed. Our proposed method is different compared to the previous works because we construct two types of hierarchies, which are the OHSUMED disease hierarchy (ODH) as our training hierarchy and the Medline abstract disease hierarchy (MADH) as testing hierarchy. Afterward, we enrich our training hierarchy by mapping the relevant features that extracted from selected categories in the OHSUMED dataset to each node or category in the ODH based on

their specific category. Next, we perform ontology alignment by employing our ontology alignment algorithm, namely "Anchor-Flood" algorithm (AFA) [16]. During ontology alignment phase, AFA matches the concepts and relations between the "enriched" OHSUMED disease hierarchy (EODH) and the MADH for exploring and searching the aligned pairs. In our research, we consider the aligned pairs as a set of probable relevant categories for classification purpose. Then, we evaluate the more specific concepts by calculating the cosine similarity score between the new Medline abstract and each probable category. Finally, we classify the new Medline abstract based on the highest cosine similarity score. In order to evaluate our approach, we conducted the experiments of the multi-class flat classification and the hierarchical classification (using the ODH and EODH). We use the results of hierarchical classification using the ODH as our baseline. The experimental evaluation indicates that our proposed approach performs slightly better than the performance of the baseline.

We organize the paper as follows: In Section 2, we describe our proposed approach to hierarchical classification. Afterward, the text preprocessing process is explained in Section 3. In Section 4 and 5, we discuss on how to construct the "OHSUMED disease hierarchy and the "enriched" OHSUMED disease hierarchy, respectively. Section 6 describes the Medline abstract disease hierarchy. The ontology alignment process is explained in Section 7. In Section 8, we discuss the ontology based hierarchical classification. Section 9 contains the experiments and in Section 10, the discussions of the classification results are stated. Finally, we conclude this paper with a summary and suggestions for future work in section 11.

2. PROPOSED APPROACH TO HIERARCHICAL CLASSIFICATION

The large number of biomedical literature that published on web makes the process of classification become challenging and arduous. The reason is that there are many categories of biomedical literature available in the web such as gene, protein, human disease, etc. and each category have many different classes. For instance, human disease category contains many different classes including heart disease, cancer, diabetes and hepatitis. Moreover, each disease class consists of many subclasses. In heart disease category contains arrhythmia, heart block, myocardial diseases, etc. Due to this situation, various classification methods were proposed by many researchers for classifying biomedical literature. For example, Nenadic et al. in [8] reported that they had employed SVM in order to classify the gene names that extracted in the molecular biology literature.

However, our approach is different from the previous works, where we explore the use of hierarchical 'concept' structure with the help of ontology alignment algorithm for searching and identifying the probable categories in order to classify biomedical text abstracts. To realize our propose method, we use the features that extracted from the datasets to index the biomedical text abstracts. These features will be used to represent our documents in order to improve the accuracy of classification performance and also the result of searching relevant documents. Due to this reason, we construct two types of hierarchies, which are the ODH as our training hierarchy and the MADH as testing hierarchy. For this purpose, initially, we construct the ODH by referring to the OHSUMED disease directory. Subsequently, we perform text preprocessing (including part-of-speech tagging, phrase chunking, etc.) for extracting and selecting the relevant features from the OHSUMED dataset and the Medline abstracts respectively. Then, we enrich the ODH by assigning the relevant features that extracted from OHSUMED dataset to each node of the hierarchy. While, the MADH were constructed using a collection of biomedical text abstracts that downloaded from the Medline database.

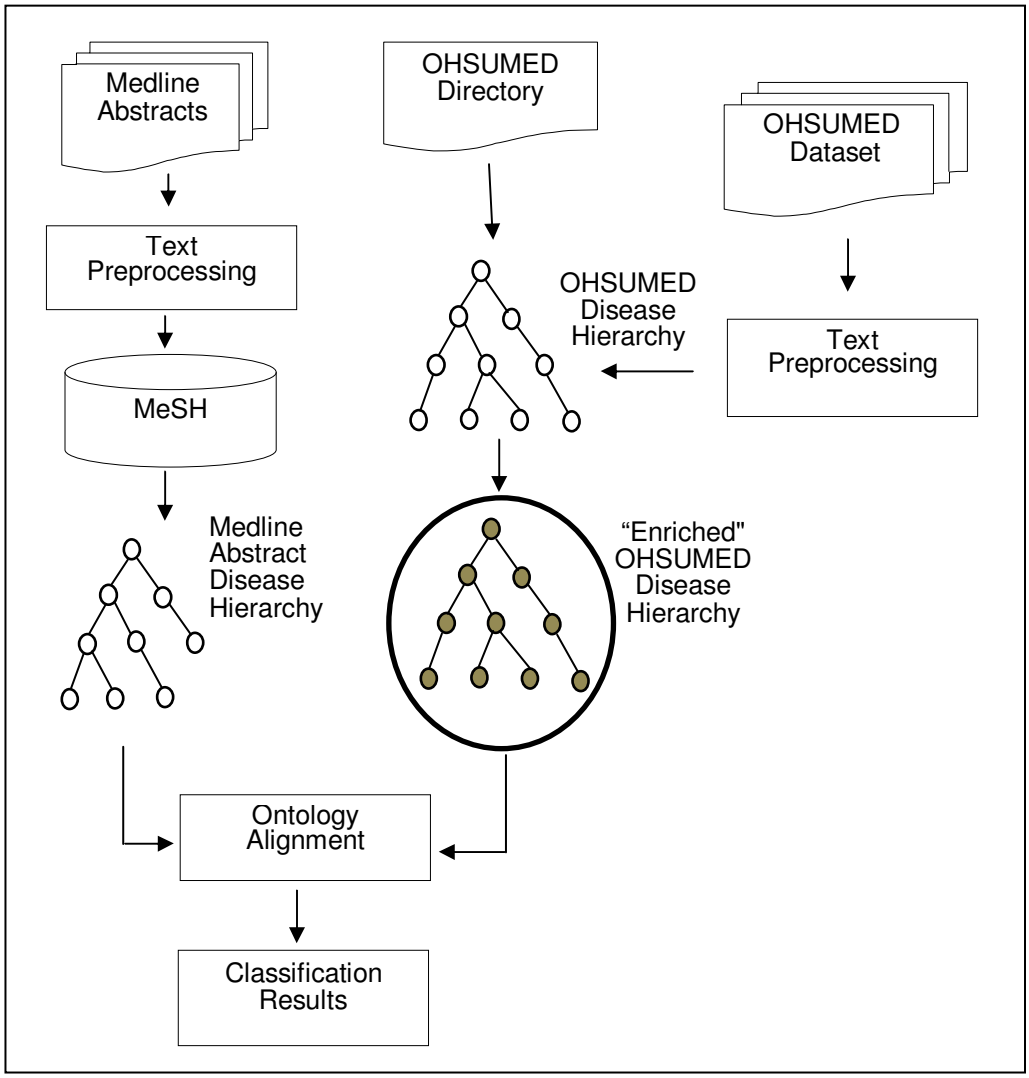

FIGURE 1: A Method for Hierarchical Classification of Biomedical Text Abstracts

Next, we perform the ontology alignment process for matching and aligning the EODH and the MADH using the "Anchor-Flood" algorithm (AFA). During ontology alignment process, AFA would match and search the concepts in both hierarchies in order to compute the similarity among the concepts and relations in the EODH and the MADH for identifying and producing the aligned pairs. We consider the aligned pairs as a set of probable categories for classifying biomedical text abstracts. Afterward, we evaluate the more specific concepts based on the cosine similarity score between the vectors of unknown new abstract in each MADH and the vectors of each probable category in the EODH for predicting more specific category. Eventually, we classify the new Medline abstracts into the first rank of cosine similarity score. Figure 1 illustrates our proposed hierarchical classification method that is implemented in our research.

3. TEXT PREPROCESSING

In our experiments, we use two different datasets, which are a subset of the OHSUMED dataset as training documents and the Medline abstracts as test documents. The OHSUMED dataset [12] is a subset of clinical paper abstracts from the Medline database, from year 1987 to 1991. This dataset contains more than 350,000 documents. However, we select 400 documents from 43 categories of the subset of OHSUMED dataset for enriching the ODH. For each category, we

retrieve on average 10 documents. For classification purpose, we have selected randomly a subset of the Medline abstracts from the Medline database. Using a PubMed website, we retrieve this dataset with the query terms, such as “arrhythmia”, "heart block", etc. to retrieve the related Medline abstracts which containing the query terms. Then, we index this dataset belonging to 15 categories of disease as shown in Table 1. A total number of Medline abstracts are 150.

In our research, each document must be represented by a set of feature vectors. Accordingly, we use noun phrases as our features. For this purpose, we perform text preprocessing to extract and select a list of unique and relevant features from our training and testing datasets. Our text preprocessing process consists of feature extraction and feature selection phase.

TABLE 1: The Number and Categories of Medline Abstracts

Category No.	Category Name	No. of Documents
1	Arrhythmia	10
2	Heart Block	10
3	Coronary Disease	10
4	Angina Pectoris	10
5	Heart Neoplasms	10
6	Heart Valve Diseases	10
7	Aortic Valve Stenosis	10
8	Myocardial Diseases	10
9	Myocarditis	10
10	Pericarditis	10
11	Tachycardia	10
12	Endocarditis	10
13	Mitral Valve Stenosis	10
14	Pulmonary Heart Disease	10
15	Rheumatic Heart Disease	10
	Total	**150**

3.1 Feature Extraction

In text preprocessing process, initially we extract the noun phrases as our features from the OHSUMED dataset and the Medline abstracts respectively. The purpose of feature extraction is to generate a list of unique features from the datasets. In our research, this process is done by performing part-of-speech (POS) tagging and phrase chunking. POS tagging can be defined as a task of assigning POS categories to terms from a predefined set of categories. Meanwhile, phrase chunking is the process of identifying and recognizing the noun phrases constructed by the POS tags. In POS tagging phase, we have automatically assigned POS tags to each term using the rule based POS tagger. Then, we extract features from the tagged text based on the chunks that consists of noun phrases. Finally, we create a list of unique features that extracted from the training and testing datasets, respectively.

3.2 Feature Selection

Feature selection phase is one of the most important tasks in text preprocessing. This is due to the fact that some features are uninformative and they do not influence the classification performance. Furthermore, as the number of unique features which extracted from our training and testing dataset is big, feature selection phase can help us to reduce the original features to a small number of features by removing the rare and irrelevant features. During feature selection phase, we attempt to find relevant features for constructing the MADH and also for enriching the ODH. Therefore, we employ the document frequency and the chi-square (χ^2) techniques to distinguish between relevant and irrelevant features, before we eliminate some percentage of the extracted features according to their document frequency and dependency between categories and features.

In order to select the relevant features for representing our datasets, we use the document frequency as a feature reduction technique for eliminating rare features. Therefore, we compute the document frequency for each unique feature in both datasets. Then, we eliminate the features with the highest and lowest frequencies. By performing this process, we could reduce the feature space into a small number of important features. Subsequently, a χ^2 test is used to measure the independence between feature (*t)* and category (*c*) in order to distinguish between relevant and irrelevant features. We attempt to identify and search the most discriminating features for each category. Then, we select the relevant features by assigning features to specific categories. We measure the relationship between features (*t)* and categories (*c*) using the following equation.

$$\chi^2 = \sum_{i,j} \frac{(o_{i,j} - e_{i,j})^2}{e_{i,j}}; \quad (1)$$

where $o_{i,j}$ is the observed frequency for each cell, while $e_{i,j}$ is the expected frequency for each cell.

In our χ^2 test, we use the 2 x 2 contingency table to compare the χ^2 distribution with one degree of freedom. Then, we rank the features according to the χ^2 score. In our experiments, we select features set by choosing features with the χ^2 score greater than 3.841 (our threshold) as the relevant features. Thereafter, we use all of selected features (as concepts) for constructing the MADH and enriching the ODH. Table 2 shows the statistic of the features used in our experiments.

TABLE 2: The Statistic of Selected Features

Experiments	Features	Number of Features
Flat Classification	Features (Before employing feature selection)	3,145
	Feature (After employing feature selection)	1,081
Hierarchical Classification	Features (Before enriching ODH)	43
	Features (After enriching ODH & before employing feature selection)	3,145
	Feature (After enriching ODH & employing feature selection)	1,081

For classification purpose, each document in our datasets is represented by a set of relevant feature vectors, whereby each feature in a vector of a document representing either a single word or multi-words in the document. Accordingly, we assign the term frequency as the feature weighting scheme to each feature for representing our documents. Term frequency $tf_{i,f}$ is the frequency of term *i* occurs in document *j* and *f* = 1, …, *m*. After text preprocessing, we assume that the document *d* is represented as follow;

$d_1 = \{ tf_{11} , tf_{12} , \dots , tf_{1m} \}$

.

.

$d_n = \{ tf_{n1} , tf_{n2} , \dots , tf_{nm} \}$

4. THE OHSUMED DISEASE HIERARCHY (ODH)

In this research, we construct the ODH by referring to the OHSUMED directory. This directory could be accessed in the OHSUMED dataset [17]. There are 101 categories in the OHSUMED directory which are divided into four levels. Level 1 contains 23 categories and level 2 consists of

56 categories. In level 3, there are 16 categories and finally, level 4 contains only 6 categories. We then construct the ODH using Protégé. Figure 2 shows the part of the OHSUMED directory, while Figure 3 illustrates the part of the ODH.

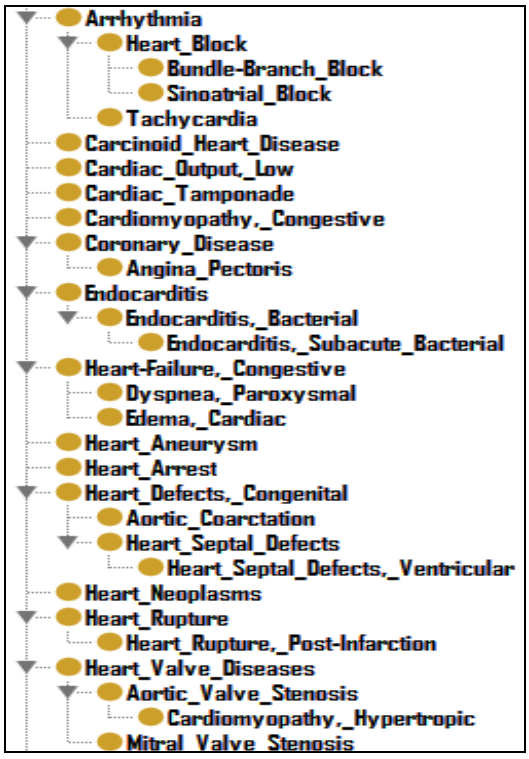

FIGURE 2: The Part of OHSUMED Directory

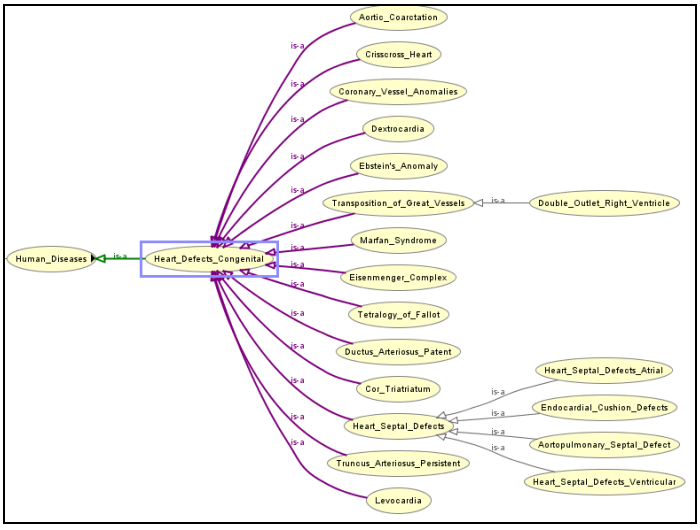

FIGURE 3: The Part of OHSUMED Disease Hierarchy

5. THE "ENRICHED" OHSUMED DISEASE HIERARCHY (EODH)

The important task in enriching the ODH is to select meaningful and relevant features from the OHSUMED dataset. In order to enrich the ODH, we select 43 categories from the OHSUMED

dataset and for each category, we retrieve about 10 documents. Afterward, we perform text preprocessing. The description of text preprocessing has been explained in Section 3.

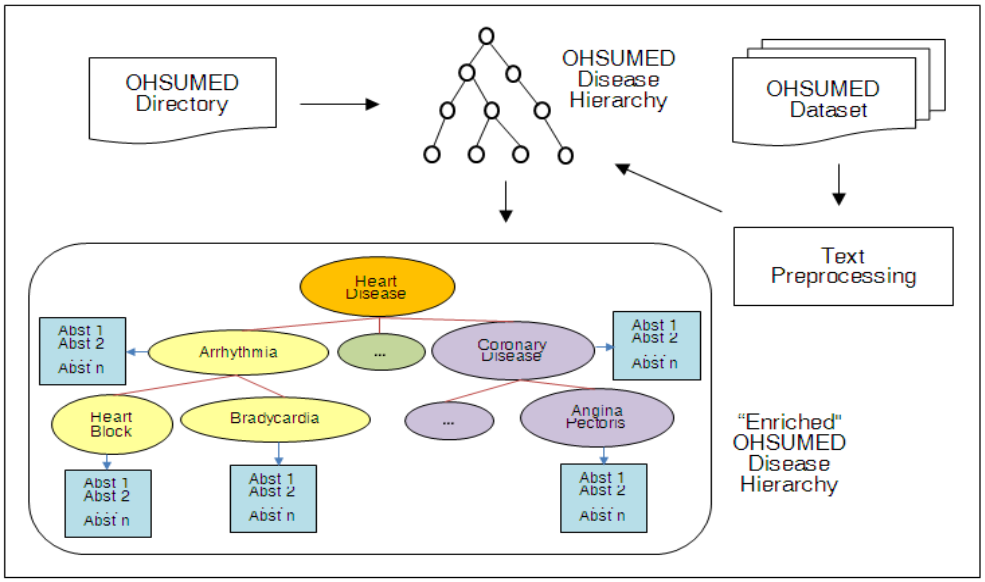

FIGURE 4: An Approach for Constructing and Enriching the ODH

During text preprocessing process, we attempt to identify and select a set of relevant features for each node. Next, we use the relevant features that extracted from 43 selected categories in the OHSUMED dataset as feature vectors. For enriching the ODH, these feature vectors are mapped to each node or concept of the ODH according to their specific category as shown in Figure 4. Meanwhile, Figure 5 describes the example of the EODH.

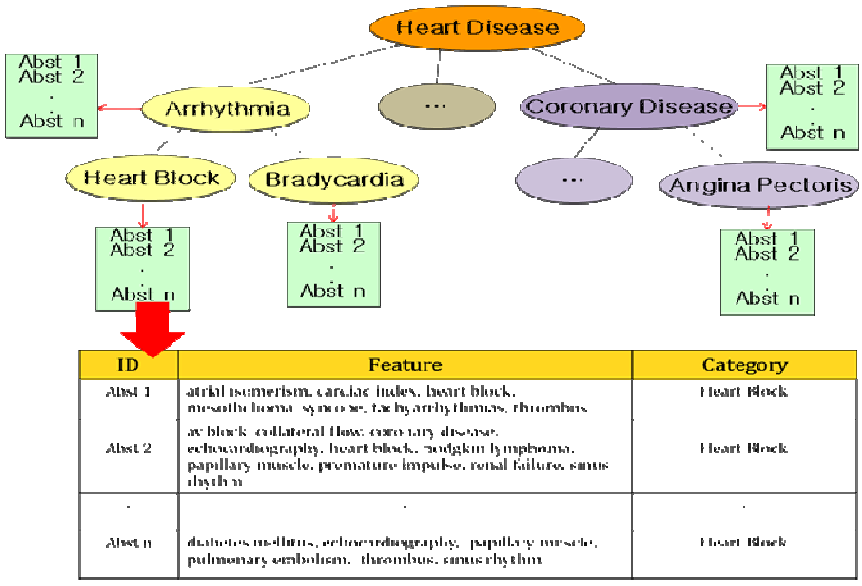

FIGURE 5: An Example of the "Enriched" OHSUMED Disease Hierarchy

6. THE MEDLINE ABSTRACT DISEASE HIERARCHY (MADH)

We construct the Medline abstract disease hierarchy using the selected features that are extracted from a collection of biomedical text abstracts that downloaded from the Medline database. Section 3 contains the description of text preprocessing process.

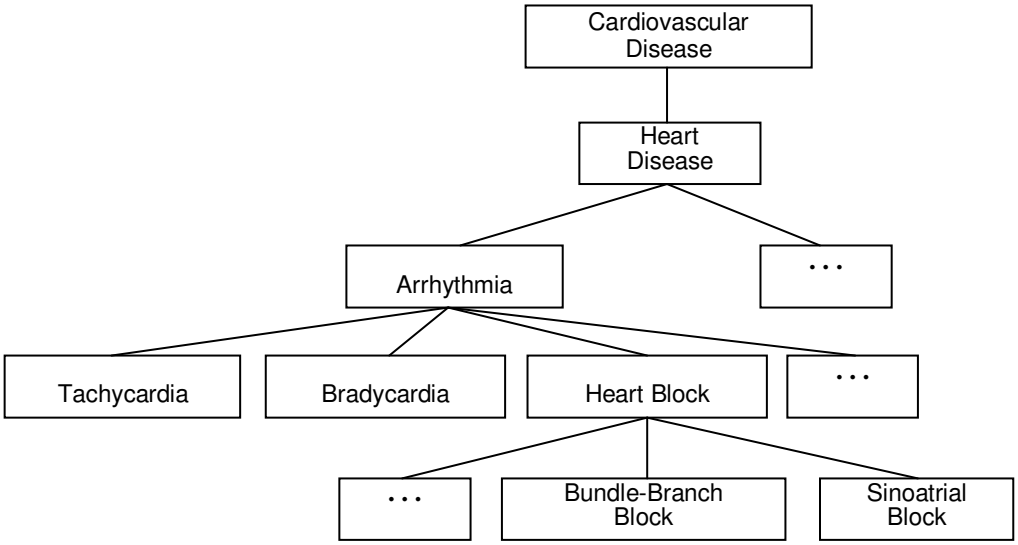

FIGURE 6: An Example of a Part of the MADH

In our research, we use Protégé for constructing the MADH. For this purpose, we have to assign the best-matching concept identifier to each selected feature. Therefore, we refer to the Medical Subject Headings (MeSH) for indexing and assigning the relevant feature in the Medline abstracts into a hierarchical structure. The reason is that the MeSH terms are arranged hierarchically and by referring to the concepts in the MeSH tree structure [18], we could identify heading and subheading of hierarchical grouping before indexing these selected features for representing our testing dataset. Finally, we construct the MADH by using the heading and subheading of hierarchical grouping that are suggested by the MeSH tree structure. Figure 6 depicts a part of the MADH.

7. ONTOLOGY ALIGNMENT

The purpose of ontology alignment in our research is to match the concepts and relations between the EODH and the MADH. Therefore, we perform ontology alignment using the "Anchor-Flood" algorithm (AFA). During ontology alignment process, AFA would explore and search for the similarity among the neighboring concepts in both hierarchies based on terminological alignment and structural alignment. Then, AFA would narrow down the EODH and the MADH for producing the aligned pairs. These aligned pairs are obtained by measuring similarity values, which consider textual contents, structure and semantics (available in the hierarchies) between pairs of entities. We consider all of the aligned pairs as a set of probable categories for classification purpose.

8. ONTOLOGY BASED HIERARCHICAL CLASSIFICATION

In our proposed approach, we construct two types of hierarchies, which are the ODH, as our training hierarchy and the MADH as testing hierarchy. Then, we perform ontology alignment in order to match both hierarchies for producing the aligned pairs, which we consider as a set of probable categories for predicting and classifying a new Medline abstract.

Afterward, we evaluate the more specific concepts based on the similarity between the new Medline abstract and each probable relevant category. Consequently, we compute the cosine

similarity score between the vector of unknown new abstract in each MADH and the vector of each probable category in the EODH for identifying and predicting more specific category. The cosine similarity score is calculated using the following equation.

$$sim(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,k}^2}} \quad (2)$$

where the vector of $d_j = (w_{11}, w_{12}, \ldots, w_{1n})$ and the vector of $d_k = (w_{21}, w_{22}, \ldots, w_{2n})$.

Then, we sort all the probable categories according to the assigned cosine similarity score. Eventually, we classify the new Medline abstracts based on the highest cosine similarity score.

9. EXPERIMENT

For our experiments, we have used 400 records (documents) from 43 different categories of the subset of the OHSUMED dataset for enriching the ODH. On the other hand, for classification purpose, we have randomly downloaded 150 biomedical text abstracts that related to human diseases, such as “arrhythmia”, "coronary disease”, “angina pectoris”, etc. from Medline database. The description of text preprocessing has been discussed in Section 3.

In order to evaluate the performance of our proposed method, we conduct two different experiments, which are multi-class flat classification and hierarchical classification. In both flat and hierarchical classification, we performed a few experiments using the features (for enriching our initial training hierarchy or ODH) that produced before and after feature selection process. Moreover, we also conducted the experiments using very few features, whereby only consists of keyword for each node in our training hierarchy (ODH) as a baseline for hierarchical classification. Then, we compare the performance of our method for classifying biomedical text abstracts with the performance of the multi-class flat classification using LIBSVM [19].

8.1 Flat Classification

In the flat classification, we have conducted the multi-class classification experiments using LIBSVM. In these experiments, we ignore hierarchical structure and treat each category separately. Then, we compare the results that produced from these experiments with the performance of our proposed method for hierarchical classification of biomedical text abstracts. We use the results of flat classification (without feature selection process) as our baseline for evaluating the performance of our proposed approach for hierarchical classification.

8.2 Hierarchical Classification

For the hierarchical classification experiments, we attempt to assign a category for a given new Medline abstract. To achieve our goal, we consider the aligned pairs were produced during ontology alignment process as a set of our probable categories. Then, we evaluate the more specific category based on the similarity between the new Medline abstract and each probable category. For this purpose, we compute the cosine similarity score between the vector of unknown new abstract in each MADH and the vector of each probable category in the EODH. After that, we sort all the probable categories according to the assigned cosine similarity score. In our research, we consider the highest cosine similarity score as the relevant category. Finally, we classify the new Medline abstracts into a category that has the highest cosine similarity score.

In order to evaluate our approach, we conduct three types of the experiments for hierarchy classification of biomedical text abstracts. Firstly, we perform the hierarchical classification using the initial training hierarchy or ODH (without enriching ODH). Then, we repeat the experiments of hierarchical classification using the “Enriched” ODH (without feature selection process). Furthermore, we also conduct the experiments of hierarchical classification using the “Enriched” ODH (with feature selection process).

10. DISCUSSION

The results of the flat and hierarchical classification are shown in Table 3, Table 4 and Figure 7. From the experiments, the results show the different performance for each category in the flat and hierarchical classification. Generally, the experimental results indicate that our proposed approach performs slightly better than the baseline. We observe that our proposed approach to hierarchical classification achieve the average accuracies of 14% (for hierarchical classification using the features in the ODH only), 30.67% (for hierarchical classification using the features in the EODH and without feature selection process) and 32.67% (for hierarchical classification using the features in the EODH and with feature selection process), respectively. Nevertheless, the accuracies of the flat classification are on the average 6.67% (for flat classification without feature selection process) and 18% (for flat classification with feature selection process), respectively.

TABLE 3: The Results of the Flat Classification

Category No.	Category Name	Flat Classification (% Accuracy)	Flat Classification + Feature Selection (% Accuracy)
1	Arrhythmia	0	20
2	Heart Block	0	0
3	Coronary Disease	0	0
4	Angina Pectoris	0	20
5	Heart Neoplasms	20	50
6	Heart Valve Diseases	10	10
7	Aortic Valve Stenosis	0	0
8	Myocardial Diseases	0	20
9	Myocarditis	10	20
10	Pericarditis	20	0
11	Tachycardia	0	30
12	Endocarditis	0	20
13	Mitral Valve Stenosis	10	30
14	Pulmonary Heart Disease	30	30
15	Rheumatic Heart Disease	0	20
Average		**6.67**	**18**

Table 3 shows the performance of the flat classification for each category of biomedical text abstracts. In general, the classification performances for category 1, 4, 5, 8, 9, 11, 12, 13 and 15 using the flat classification approach (with feature selection process) are better than the flat classification approach (without feature selection process). For instance, the performance of flat classification approach (with feature selection process) reach the highest accuracy (50%) in category 5, while the classification accuracy of the flat classification approach (without feature selection process) only achieve 20% in the same category, as shown in Table 3. These results might demonstrate that if the relevant features are selected carefully for representing a document, the accuracy of biomedical text abstracts classification would be increased.

In addition, we have compared the performances of hierarchical classification using different approaches. Table 4 illustrates the effect of enriching the ODH and employing the feature selection process for classifying 15 categories of biomedical text abstracts. For the hierarchical classification approach (using EODH and with feature selection process), the results of the category 1, 5, 7, 8, 10, 12, 13, 14 and 15 show the best performance compared to the hierarchical classification approach (using EODH and without feature selection process) and the hierarchical classification approach (using ODH only) as shown in Table 4.

TABLE 4: The Results of the Hierarchical Classification

Category No.	Category Name	Hierarchical Classification + ODH (% Accuracy)	Hierarchical Classification + "Enriched" ODH (% Accuracy)	Hierarchical Classification + "Enriched" ODH + Feature Selection (% Accuracy)
1	Arrhythmia	0	50	50
2	Heart Block	0	0	0
3	Coronary Disease	0	0	0
4	Angina Pectoris	0	0	0
5	Heart Neoplasms	0	30	30
6	Heart Valve Diseases	0	0	0
7	Aortic Valve Stenosis	10	20	20
8	Myocardial Diseases	0	40	50
9	Myocarditis	0	0	0
10	Pericarditis	40	60	70
11	Tachycardia	0	0	0
12	Endocarditis	40	70	70
13	Mitral Valve Stenosis	10	60	70
14	Pulmonary Heart Disease	20	40	40
15	Rheumatic Heart Disease	90	90	90
Average		**14**	**30.67**	**32.67**

Furthermore, the classification accuracies for the hierarchical classification approach (using EODH and without feature selection process) produce quite similar results to the performance of hierarchical approach (using EODH and with feature selection process) for all categories except for the category 8 (40%), category 10 (60%) and category 13 (60%). Nonetheless, for the hierarchical approach (using ODH), the classification accuracies of the some categories such as in category 10 and 15 achieve quite good results. Overall, the results for the category 15 using the hierarchical classification approaches are better compared to other categories, which achieve 90% of accuracy. These results indicate that the hierarchical classification approaches has been able to classify the biomedical text abstracts correctly although the number of training documents representing each category is small.

By comparing Table 3 and Table 4, we observe that overall average accuracies of the hierarchical classification show better performance than the average accuracies of the flat classification. For example, the classification accuracies for the category 12 and 15 show better results when employing the hierarchical classification approaches compared with the flat classification approaches as shown in Table 3 and Table 4. According to the results that are obtained from the experiments, we can say that the hierarchical classification approaches can increase the classification accuracy because our proposed approach can identify and propose more relevant category for classifying the biomedical text abstracts by using our ontology alignment algorithm.

Additionally, we also noticed that the feature selection process has little influence on the classification performance in both flat and hierarchical classification. The results of the flat classification approach (with feature selection process) show quite good performance than the flat classification approach (without feature selection process), especially for category 5 and 11, as shown in Table 3. On the other hand, the classification performance of hierarchical classification approach (with feature selection process) is slightly better than the classification performance of hierarchical classification approach (without feature selection process) as shown in the Table 4 and Figure 7. The average accuracies of the flat and hierarchical classification approaches (with feature selection process) are 18% and 32.67%, respectively. These results indicate that the flat and hierarchical classification approaches (with feature selection process) has been able to classify a new unknown biomedical text abstract correctly even though we use a small number of relevant features for representing each category.

In addition, the classification accuracies of a few categories such as in category 12 and 13 produce better results in both flat and hierarchical classification (with feature selection process) experiments compared to other diseases categories. This might be caused by the number of content-bearing keywords that are extracted and selected in these categories are higher than other categories. Moreover, the results of the flat and hierarchical classification for some categories such as category 2 and 3 show 0% accuracies. The main reason of the performances of flat and hierarchical classification for these categories being poor is that the selected features for representing each document are sparse or common features.

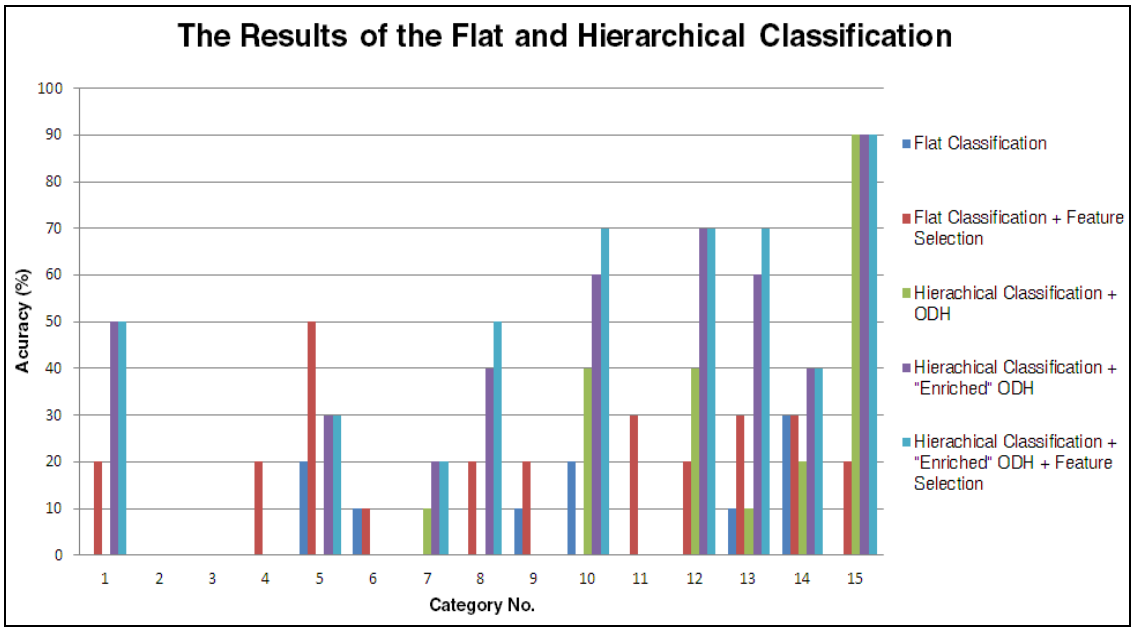

FIGURE 7: The Performance of Classification Accuracies

Although the performance of hierarchical classification experiments produced better results than the flat classification, our proposed approach is still naïve in achieving the good classification accuracies. The low performance in the hierarchical classification might be caused by the shortness of the Medline abstracts or extraction of very few relevant features. Consequently, we construct a small hierarchy for ontology alignment purpose, which may produce a small number of aligned pairs (as our probable category). Furthermore, the small number of documents that are represented in a particular category in the dataset may also affect the decrease of the classification accuracy. Even though the performance of our proposed approach for hierarchical classification is still below than our target, we believe that we can improve the classification accuracies. We are confident that, by enriching the ODH and also the MADH with the relevant features, we can identify more probable categories for classifying biomedical text abstracts with the help of our ontology alignment algorithm.

11. CONCLUSION

The main purpose of our research is to improve the performance of hierarchical classification by increasing the accuracies of classes in the datasets that are represented with a small number of biomedical text abstracts. Therefore, in this paper, we propose the hierarchical classification approach that utilizes the ontology alignment algorithm for classification purpose. Our approach is different from the previous works, where we explore the use of hierarchical 'concept' structure with the help of our ontology alignment algorithm, namely 'Anchor-Flood' algorithm (AFA) for searching and identifying the probable categories in order to classify the given biomedical text abstracts.

Then, we evaluate the performance of our approach by conducting the hierarchical classification experiments using the features that are extracted from the OHSUMED dataset and Medline abstracts. We also conduct the multi-class flat classification experiments using LIBSVM. Moreover, we perform feature selection by employing the document frequency and chi-square techniques for selecting relevant features. Then, we perform a few experiments for the flat and hierarchical classification using the features that are selected from feature selection process. Generally, the experimental results indicate that our propose approach of hierarchical 'concept' structure using ontology alignment algorithm can improve the classification performance. Although our proposed approach is still naïve in achieving the good classification accuracies, we believe that we could modify our proposed approach to produce more relevant probable categories and predict more specific category for classifying biomedical text abstracts.

Our future target is to seek and propose more accurate approaches for selecting relevant and meaningful features in order to enrich or expand the ODH and MADH. These features would be used to index the biomedical text abstracts for increasing the accuracy of classification performance and also the result of searching relevant documents. Furthermore, the performance of our proposed approach for hierarchical text classification also can be improved by increasing the total number of documents that are represented in each category in the dataset.

ACKNOLEDGEMENT

This work was supported in part by Global COE Program "Frontiers of Intelligent Sensing" from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

12. REFERENCES

1. A. M. Cohen. *"An effective general purpose approach for automated biomedical document classification"*. AMIA Annual Symposium Proceeding, 2006:161-165, 2006

2. A. Sun and E. Lim. *"Hierarchical text classification and evaluation"*. In Proceeding of the IEEE International Conference on Data Mining. Washington DC, USA, 2001

3. F. M. Couto, B. Martins and M. J. Silva. *"Classifying biological articles using web sources"*. In Proceedings of the ACM Symposium on Applied Computing. Nicosia, Cyprus, 2004

4. 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. Las Vegas, NV, 2005

5. A. Pulijala and S. Gauch. *"Hierarchical text classification"*. In Proceedings of the International Conference on Cybernetics and Information Technologies (CITSA). Orlando, FL, 2004

6. S. Gauch, A. Chandramouli and S. Ranganathan. *"Training a hierarchical classifier using inter-document relationships"*. Technical Report, ITTC-FY2007-TR-31020-01, August 2006

7. M. E. Ruiz and P. Srinivasan. *"Hierarchical text categorization using neural networks"*. Information Retrieval, 5(1):87-118, 2002

8. T. Li, S. Zhu and M. Ogihara. *"Hierarchical document classification using automatically generated hierarchy"*. Journal of Intelligent Information Systems, 29(2):211-230, 2007

9. K. Deschacht and M. F. Moens. *"Efficient hierarchical entity classifier using conditional random fields"*. In Proceedings of the 2nd Workshop on Ontology Learning and Population. Sydney, Australia, 2006

10. G. R. Xue, D. Xing, Q. Yang and Y. Yu. *"Deep classification in large-scale text hierarchies"*. In Proceeding of the 31st Annual International ACM SIGIR Conference. Singapore, 2008

11. G. Nenadic, S. Rice, I. Spasic, S. Ananiadou and B. Stapley. *"Selecting text features for gene name classification: from documents to terms"*. In Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine, PA, USA, 2003

12. Y. Wang and Z. Gong. *"Hierarchical classification of web pages using support vector machine"*. Lecture Notes in Computer Science, Springer, 5362/2008:12-21, 2008

13. S. Dumais and H. Chen. *"Hierarchical classification of web content"*. In Proceedings of 23rd ACM International Conference on Research and Development in Information Retrieval. Athens, Greece, 2000

14. T. Y. Liu, Y. Yang, H. Wan, H. J. Zeng, Z. Chen and W. Y. Ma. *"Support vector machines classification with a very large-scale taxonomy"*. ACM SIGKDD Explorations Newsletter – Natural language processing and text mining, 7(1):36-43, 2005

15. G. Nenadic and S. Ananiadou. *"Mining semantically related terms from biomedical literature"*. Journal of ACM Transactions on Asian Language Information Processing, 5(1):22-43, 2006

16. M.H. Seddiqui and M. Aono. *"An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size"*. Web Semantics: Science, Services and Agents on the World Wide Web, 7:344-356, 2009

17. OHSUMED dataset. Dataset available at http://davis.wpi.edu/xmdv/datasets/ohsumed.html, 2005

18. Medical Subject Heading (MeSH) tree structures. Available at http://www.nlm.nih.gov/mesh/trees.html, 2010

19. C.-C. Chang and C.-J. Lin. *"LIBSVM: a library for support vector machines"*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2007