Employing Ontology Enrichment Algorithm in Classifying Biomedical Text Abstracts

Rozilawati binti Dollah[1,2] and Masaki Aono[1]

[1] Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi, 441-8580 Japan

[2] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

Abstract

The application of text classification systems on biomedical literature aims to select articles relevant to a specific issue from large corpora. As the amount of online biomedical literature grows, the task of finding relevant information becomes very complicated, due to the difficulties in browsing and searching the relevant information through the web. Ontology is useful for organizing and navigating the Web sites and also for improving the accuracy of Web searches. It provides a shared understanding of domain, to overcome differences in terminology such as synonym, term variants and terms ambiguity. However, one of the problems raised in ontology is the maintenance of these bases of concepts. Therefore, we investigate and propose an ontology enrichment algorithm as one of the methods to modify an existing ontology. In this research, we present a new ontology enrichment algorithm for assigning or associating each concept in the training ontology with the relevant and informative features from biomedical information sources. Experiments are conducted to extract and select the meaningful features from different information sources such as the OHSUMED dataset, Medical Subject Heading (MeSH) terms and heart disease glossaries. Then, we expand these features into the training ontology. Finally, we evaluate the performance of our proposed ontology enrichment algorithm in classifying biomedical text abstracts. The results demonstrate that the macro-average for precision, recall and F measure are improved by employing ontology enrichment algorithm.

Keywords: *MeSH, OHSUMED, Ontology Enrichment, Text Classification, Text Mining.*

1. Introduction

As the number of online biomedical literature grows, the task of finding relevant information becomes very complicated. This is due to the fact that, it will become much more difficult to browse and search the relevant information from the web. The overwhelming amount of published biomedical knowledge in texts, demands for effective automated processing that can help researchers to gather and make use of the knowledge. Nowadays, most search engines rely on subject indexing system compared with document or text classification, in order to find relevant articles that researchers required. Subject indexing system based on keyword is capable of finding articles that contain specific keyword using exact matching. However, it is difficult to identify the ambiguous medical terms collected in biomedical articles. This is because many biological terms and their variant are ambiguous.

In addition, lack of standardized biomedical vocabulary for indexing the biomedical articles also causes difficulties for the researchers to effectively and efficiently organize and retrieve relevant articles from the web such as PubMed. PubMed is a free online database developed by the US National Library of Medicine (NLM) that provides access to the database [1]. It has over 23 million abstracts and references on biomedical literature from MEDLINE and life science journals. For that reason, organizing and classifying biomedical articles is a challenging task [2], especially when a large number of biomedical articles should be organized into a hierarchical structure or ontology. Ontology consists of list of terms or features and the relationships between these terms or features. It provides a shared understanding of domain, to overcome differences in terminology such as synonym, term variants and terms ambiguity.

According to Antoniou and Harmelen [3], ontology is useful for organizing and navigating the Web sites and for improving the accuracy of Web searches. Ontology also can be used to standardize terminology, to enable access to domain knowledge, to verify data consistency and to facilitate integrative analyses over heterogeneous biomedical literature. However, one of the problems in ontology research is the maintenance of the existing ontology. Due to this problem, ontology enrichment could be one of solution in order to modify the existing ontology.

Recently, research on ontology enrichment area has received broad attention and various ontology enrichment approaches or methods were proposed by many researchers to overcome the problem of the lack of coverage of ontologies for improving the results of classification accuracy and retrieving relevant information. Even though they achieved good performance in precision, there are still some issues in achieving good recall.

Due to this problem, we attempt to propose and employ a new ontology enrichment algorithm to enhance the relevant features into the training ontology in order to predict more specific categories for classifying biomedical literature which would increase the performance of classification method. In our proposed algorithm, we assign or expand more relevant and informative features that represent each node or concept of training ontology. For this purpose, we extract and select the relevant features from different biomedical information sources such as from a collection of biomedical text abstracts from a subset of the OHSUMED dataset, MeSH entry terms and heart disease glossaries. Finally, we conduct experiments to evaluate the effectiveness of the proposed ontology enrichment algorithm using these biomedical information sources.

We organize the paper as follows. In Section 2, we summarize a review of earlier work. Section 3 describes the hierarchical classification approach. Then, our proposed ontology enrichment algorithm employed in the experiments is described in Section 4. In addition, Section 5 explains the discussion of the experiments and results. Finally, Section 6 concludes and summarizes the paper and suggests directions for future work related to this research.

2. Related Work

According to Geeta [4], due to the overwhelming amount of online free format text and the difficulties of retrieving relevant information, there is an escalating need for proper organization of the available data in an efficient structure. She argued that classification is the first step towards organization in order to distinguish the data and assign the relevant data under an appropriate category. In addition, the large amount of online data, especially in biomedical literature that has been published on the web through sites such as PubMed, makes the process of classification challenging and arduous. This is due to the many number of categories of biomedical literature available in the PubMed database such as gene, protein, human disease and each different category has many different classes. For instance, the human disease category contains many different classes including heart disease, cancer, diabetes and hepatitis. Moreover, each disease class consists of many subclasses. For example, the heart disease class contains subclasses arrhythmia, myocardial diseases and so forth.

To improve classification accuracy and retrieve more relevant information, various classification methods have been proposed by many researchers, including the hierarchical classification method. For example, Pulijala and Gauch [5] and Gauch et al. [6] have classified the documents during indexing which can be retrieved by using a combination of keyword and conceptual match. Li, et al. [7] proposed another approach of hierarchical document classification using linear discriminant projection to generate topic hierarchies.

Furthermore, Deschacht and Moens [8] proposed an automatic hierarchical entity classifier for tagging noun phrases in a text with their WordNet synset using conditional random fields. Meanwhile, Xue, et al. [9] developed a deep-classification algorithm to classify web documents into categories in large-scale text hierarchy. In addition, Antoniou and Harmelen [3] stated that, ontologies are useful for organizing and navigating the Web sites and also for improving the accuracy of Web searches. However, the maintenance of the existing ontology becomes one of the problems in ontology research area. One of the solutions in modifying the existing ontology is by employing the ontology enrichment approach.

Recently, many ontology enrichment approaches were proposed to expand the relevant features for improving the results of classification accuracy and retrieving more relevant information. Some researchers such as in [10] and [11] have tried to extract words from the WordNet lexical enrich the vocabularies. In other research, Arnold and Rahm [12], proposed a new semantic ontology matching by implementing enrichment approach to improve the mapping by identifying several is-a and inverse is-a relationship. Although many hierarchical classification methods and ontology enrichment approaches have been proposed in various domains in recent years such as in [4]-[16], there has been limited focus on the hierarchical classification and ontology enrichment of biomedical literature. Due to this predicament, this research investigates the application of ontology enrichment algorithm to associate the informative and meaningful features to represent each concept or category of 'training ontology' in order to increase the performance of our hierarchical classification.

3. Hierarchical Classification Approach

To realize our method, we have constructed two types of hierarchies or ontologies, which are training and testing ontologies. For this purpose, initially we constructed training ontology by referring to the OHSUMED directory in the OHSUMED dataset [17]. Meanwhile, testing ontology was constructed using biomedical text abstract that randomly selected from a subset of the OHSUMED dataset. Consequently, we performed the ontology alignment process to match and align the concepts in the training and testing ontologies for searching the category candidates or probable concepts using the 'Anchor-Flood' algorithm (AFA) [18].

During ontology alignment process, AFA matched and searched the concepts in both hierarchies in order to compute the similarity among the concepts and relations in training ontology and testing ontology for identifying and producing the aligned pairs. We considered the aligned pairs as a set of probable categories for classifying biomedical text abstracts. Afterward, we evaluated the more specific concepts based on the cosine similarity score between the vectors of each testing ontology and the vectors of each probable category in the 'enriched' training ontology for predicting more specific category. Eventually, we classified each testing document into the first rank of cosine similarity score.

3.1 Data Preparation

The process of data preparation involves text gathering and text preprocessing. In this research, we use 7,500 biomedical text abstracts from a subset of the OHSUMED dataset for enriching the initial training ontology, while another 500 biomedical text abstracts from OHSUMED dataset are used as testing documents. In addition, we also collect MeSH entry terms and heart disease glossaries for enriching purpose.

The purpose of text preprocessing is to extract and select noun phrases as features from these biomedical information sources. A feature represents a single or multi-words term having unique meaning, while a relevant feature is an informative word or term that would influence the classification performance. Text preprocessing process consists of feature extraction and feature selection phases. Feature extraction phase is one of the most important tasks in text preprocessing which involves part-of-speech (POS) tagging and phrase chunking. As the number of unique features which are extracted from a subset of OHSUMED dataset and heart disease glossaries is large, the feature selection phase can help us to reduce the original features to a small number of features by eliminating some features which are uninformative and do not influence the classification performance.

In this research, we eliminate the rare features by employing the document frequency technique. While, χ^2 is used to measure the independence between feature (t) and category (c) in order to distinguish between relevant and irrelevant features. We choose χ^2 for selecting the relevant features because χ^2 is a normalized value. In addition, χ^2 values are comparable across features or terms for the same category compared to other feature selection techniques such as mutual information. We attempt to identify and search for 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 in the contingency table, while $e_{i,j}$ is the expected frequency for each cell in the contingency table.

In the χ^2 test, we use the 2 x 2 contingency table to compare the χ^2 distribution with one degree of freedom. We then select the features set by choosing features with χ^2 score greater than 3.841 (our threshold) as the relevant features. Finally, we create a list of unique and relevant features that are extracted from different biomedical information sources for enriching the initial training ontology.

3.2 Construct Training Ontology

Initially, the training ontology is constructed by referring to the OHSUMED directory that can be accessed in the OHSUMED dataset [17]. There are 101 categories of heart disease in the OHSUMED directory which are divided into four levels. Level 1 is the upper level containing 23 categories and level 2 consists of 56 categories. In level 3, there are 16 categories and level 4 contains only 6 categories. We consider all 101 categories of heart disease in the OHSUMED directory as concepts or nodes for representing the initial training ontology. Afterwards, we construct the initial training ontology using Protégé. Fig. 1 shows the part of training ontology.

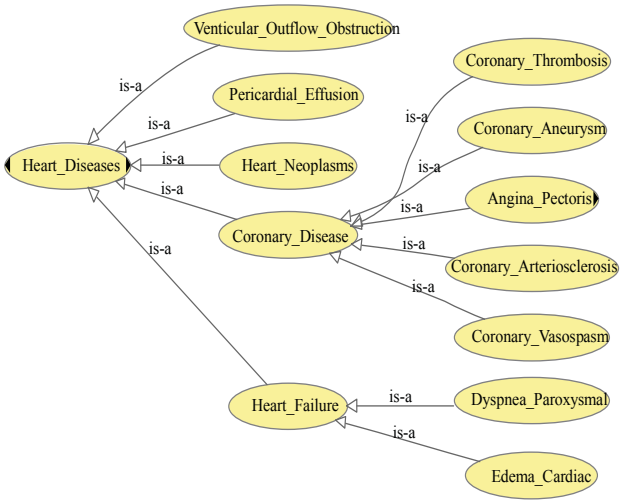

Fig. 1 The part of training ontology.

3.3 Enrich Training Ontology

The aim of this phase is to enhance the initial training ontology for improving the performance of classification accuracy. Each concept or node in the initial training ontology represents one category of heart disease. During the ontology enrichment process, each concept or category of training ontology is associated with the relevant or informative features. Therefore, we propose a new ontology enrichment algorithm to enrich the initial training ontology. Then, we employ this algorithm by assigning the relevant features that are selected from different biomedical information sources such as a collection of biomedical text abstracts (from a subset of the OHSUMED dataset), MeSH entry terms and heart disease glossaries or dictionaries. OHSUMED dataset is a subset of biomedical paper abstracts from the Medline database, from year 1987 to 1991. This dataset contains more than 350,000 documents [17]. However, we have randomly selected about 7,500 biomedical text abstracts (from a subset of the OHSUMED dataset) to enrich the initial training ontology.

For the experiments, each document must be represented by a set of feature vectors. For each biomedical information sources, we perform text preprocessing in order to extract and select noun phrases as the relevant features. The description of text preprocessing has been explained in Subsection 3.1. Then, we identify and assign the most discriminating features for each node or category of the initial training ontology.

3.4 Construct Testing Ontology

In this research, we construct testing ontology using biomedical text abstract that are randomly collected from the OHSUMED dataset. We collect 500 biomedical text abstracts (different biomedical text abstracts that used for enriching the training ontology) from the OHSUMED dataset as testing documents. For the experiments, each testing document must be represented in hierarchy or ontology. We perform text preprocessing to extract and select a list of unique and relevant features from the testing documents. Subsection 3.1 contains the description of text preprocessing.

In our research, we create testing ontology for each of testing document using Protégé. For this purpose, we refer to the Medical Subject Headings (MeSH) tree structure [19] for indexing and assigning the selected features from testing documents into a hierarchical structure. Next, we map each of these features to the concepts in the MeSH tree structure to identify the heading and subheading of hierarchical grouping. Finally, we construct the testing ontology using the heading and subheading of hierarchical grouping that are referred from the MeSH tree structure.

3.5 Perform Ontology Alignment

During the ontology alignment process, both “enriched” training and testing ontologies would be matched and aligned using the “Anchor-Flood” algorithm (AFA). The AFA would compute the similarity between concepts and relations in both ontologies for identifying and collecting the aligned pairs. We choose the AFA for matching the “enriched” training and testing ontologies as it can align ontologies of arbitrary size and search the aligned pairs across ontologies based on the structural and terminological similarity scores among the concepts and relations in both ontologies. In our experiments, we consider the aligned pairs as a set of probable categories for classifying testing documents.

3.6 Calculate Similarity Measure

We evaluate the more specific concepts based on the cosine similarity score between the feature vectors of each testing ontology and the feature vectors of each particular probable category in the training ontology for predicting more specific category. We repeatedly compute the similarity score for all probable categories. Finally, we classify each testing document into the first rank of cosine similarity score.

3.7 Evaluation

In the testing phase, we test and evaluate the effectiveness of employing the ontology enrichment algorithm for classifying biomedical text abstracts. For this purpose, we assign the relevant features that selected from different biomedical information sources such as the biomedical text abstracts (OHSUMED dataset), MeSH entry terms, heart disease glossaries and also a combination of features from MeSH entry terms and heart disease glossaries. Then, we compute the precision, recall and F measure for each experiment for evaluating the performance of the proposed ontology enrichment algorithm.

4. Ontology Enrichment Algorithm

We attempt to increase the performance of the hierarchical classification method for classifying biomedical text abstracts. Consequently, we investigate and explore the application of ontology enrichment in our hierarchical classification method [13], [14]. We are confident that ontology enrichment is one of the most important steps for constructing the training ontology. During the ontology enrichment process, each category or concept of training ontology is associated with the relevant or informative features. If the more relevant features are added into each category, the possibility of that category being categorized becomes higher. In this research, we attempt to organize and associate more meaningful and relevant features to represent each category or concept in the initial training ontology by employing our proposed ontology enrichment algorithm.

Towards this effort, we extract and select the relevant features from different biomedical information sources such as biomedical text abstracts (OHSUMED dataset), MeSH entry terms and 12 heart disease glossaries [20]-[31] to the specific category or concept in training ontology. We repeatedly enrich the initial training ontology by considering noun phrases from these biomedical information sources. In addition, we use a combination of features from MeSH entry terms and heart disease glossaries.

Fig. 2 illustrates a flow of the proposed ontology enrichment algorithm that employed in the hierarchical classification method. In this paper, we identify 80 leaf nodes in the initial training ontology. Each leaf node is assigned a set of relevant features representing the concept or category. For each child node, the relevant features from their leaf node would be merged and normalized in order to eliminate the redundant features. The enrichment process would persist until all nodes in the training ontology have their own set of features. We repeatedly employ this algorithm for enriching the initial training ontology using the selected features from the biomedical text abstracts (OHSUMED dataset), MeSH entry terms, heart disease glossary and also a combination of features from the MeSH entry terms and heart disease glossaries. Next, we count term frequency as the feature weighting scheme for selected features from the OHSUMED dataset, MeSH tree structures and heart disease glossary. In contrast, we assign different coefficients to calculate the feature weight for a combination of features from MeSH entry terms and heart disease glossaries.

Algorithm 1 Ontology Enrichment

Input: 1) An initial training ontology
2) A set of relevant features (extracted from different biomedical information sources) for each leaf node

```
BEGIN
Initialize each node (in training ontology) as empty
REPEAT
  FOR each node i = 1, ..., N DO
  BEGIN
    IF i is a leaf node THEN
      Assign the relevant features to the leaf node
    ELSE
      Merge the relevant features from the child node(s)
      Normalize the redundant features
      Save the features in current node
    END IF
  END FOR
UNTIL all nodes in training ontology have their set of relevant features
END Algorithm
```

Output: An enrichment ontology

Fig. 2 The proposed ontology enrichment algorithm.

4.1 Ontology Enrichment using OHSUMED Dataset

We enrich each concept in the training ontology with the relevant noun phrases extracted from biomedical text abstracts (a subset of the OHSUMED dataset). The OHSUMED dataset [16] is a subset of clinical paper abstracts from the MEDLINE database, from 1987 to 1991. In this research, we select 7,500 biomedical text abstracts from 101 categories of the subset of the OHSUMED dataset for enriching the training ontology. Initially, we perform text preprocessing for biomedical text abstracts from a subset of the OHSUMED dataset. We then select the highly-related features from this dataset by employing the document frequency and chi-square techniques. For this purpose, we distinguish the relevant features and noise features by ranking the features based

on their chi-square score. Next, we generate a set of features by choosing features with a chi-square score greater than our threshold as the relevant features. As a result, we retrieve about 7,565 features from this dataset. These features would be used to expand the concepts in the training ontology.

4.2 Ontology Enrichment using MeSH Entry Terms

We attempt to reduce the problem of disambiguation by considering the expanded features based on the MeSH entry terms. Therefore, we map each leaf node in the initial training ontology to the MeSH tree structures to identify and collect all relevant terms to be added to each leaf node. We consider the synonyms and term variants that are collected from the MeSH tree structures as the relevant features. Subsequently, we create a list of MeSH entry terms for each leaf node in the training ontology. We retrieve about 892 features that are related to each leaf node in the training ontology. Finally, we employ these features to expand our training ontology.

4.3Ontology Enrichment using Heart Disease Glossaries

We also attempt to enrich each leaf node in the training ontology with the terms or features extracted from heart disease glossaries. Consequently, we select and collect the important features from 12 heart disease glossaries [20]-[31]. As a result, we collect about 1,362 features from 12 heart disease glossaries. Eventually, we assign these features to the related lead node in training ontology.

4.4Ontology Enrichment using Combination of MeSH Entry Terms and Heart Disease Glossary

We also investigate the effectiveness of our proposed ontology enrichment algorithm by employing the features from the MeSH tree structures and heart disease glossaries. In this paper, we assign different coefficients for calculating the feature weight for all features that appear in MeSH tree structures, heart disease glossaries and both biomedical information sources. We collect about 2,132 unique features from a combination of MeSH entry terms and heart disease glossary terms. Then, we attempt to optimize these features by assigning different coefficients to the term frequency for each feature that appears in the MeSH tree structures, heart disease glossaries and both biomedical sources. Next, we calculate the feature weight using the equation expressed below. Finally, we employ these features for expanding the initial training ontology.

$$w_k = \sum_{i,f} \mu_{i,f} (tf_{i,f}) \quad (2)$$

where $\mu_{i,j}$ is the coefficient assigned for each feature, while $tf_{i,f}$ is the feature or frequency for each term or feature

Fig. 3 shows the performance of hierarchical classification using different coefficients assigned to the selected features in the hierarchical classification experiments. Based on our observation, the features that assigned with coefficient 1,1,5 achieve top-notch performance. For this experiment, we assign coefficient 1 for each feature that appearing either only in MeSH tree structures or heart disease glossaries, while we assign coefficient 5 to each feature that appearing in both MeSH tree structures and heart disease glossaries. This performance may show that the features appearing in both the MeSH tree structures and heart disease glossaries are more important than the features appearing either only in the MeSH tree structures or heart disease glossaries.

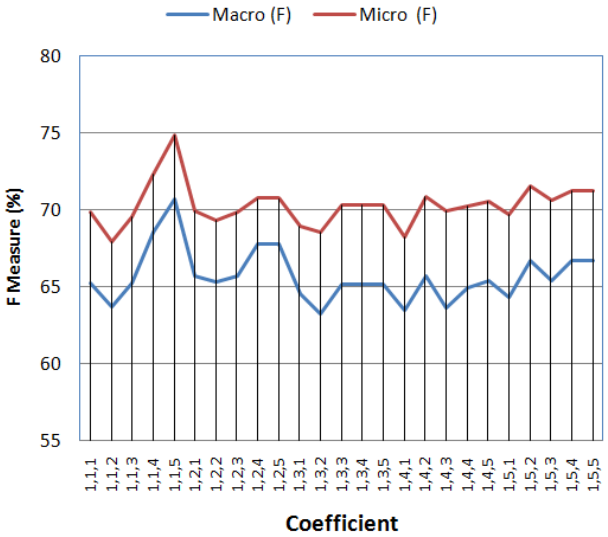

Fig. 3 Results of hierarchical classification using different coefficients for features in different biomedical information sources (MeSH, Glossary, MeSH + Glossary).

5. Experiments and Results

In this paper, we attempt to evaluate the performance of our proposed ontology enrichment algorithm in classifying biomedical text abstracts. For this purpose, we propose a new ontology enrichment algorithm for assigning or associating each concept in the initial training ontology with the relevant and informative features that are extracted and selected from different biomedical

information sources, such as biomedical text abstracts (OHSUMED dataset), MeSH entry terms, heart disease glossaries and a combination of features from MeSH entry terms and heart disease glossaries. Therefore, we conduct five different experiments for employing and evaluating the performance of our proposed ontology enrichment algorithm, as follows:

i. Experiment I - classification without ontology enrichment.
ii. Experiment II - classification with ontology enrichment using the OHSUMED dataset.
iii. Experiment III - classification with ontology enrichment using MeSH entry terms (as baseline).
iv. Experiment IV - classification with ontology enrichment using heart disease glossaries.
v. Experiment V - classification with ontology enrichment using combinations of features from MeSH entry terms and heart disease glossary.

Table 1 shows the number of selected features for different biomedical information sources used as feature vectors in the experiments. During the training phase, sets of feature vectors belonging to each category are used to represent each category. While in the classification phase, these sets of feature vectors and also the features appearing in testing documents would be used to calculate the similarity scores before assigning the relevant category.

Table 1: Number of features for each classification experiment

Experiment	No. of features
Without Enrichment	101
OHSUMED	7,565
MeSH (Baseline)	892
Glossary	1,362
MeSH + Glossary	2,132

In the hierarchical classification experiments, we attempt to assign a category for each testing document. To achieve our goal, we perform ontology alignment to match and align both the training and testing ontologies for producing the aligned pairs. In the experiments, we consider these aligned pairs as a set of probable categories for predicting and classifying each testing document. We then compute the cosine similarity score between the vector of each testing document and the vector of each probable category in the "enriched" training ontology for identifying and predicting more specific categories. The cosine similarity score is defined as follows:

$$sim(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,k}^2}} \quad (3)$$

Next, we sort all probable categories according to the assigned cosine similarity score. In this paper, we consider the category with the highest cosine similarity score as the relevant category and we assign the testing documents into this relevant category. We repeatedly run the hierarchical classification experiments by employing the ontology enrichment algorithm using different biomedical information sources. We evaluate the performance of the ontology enrichment algorithm for classifying biomedical text abstracts. Initially, we compute the precision, recall and F measure for each category. Then, we compare the performance of each hierarchical classification experiments by calculating the macro-average for precision, recall and F measure for the categories as defined as follows:

$$\hat{P}^M = \frac{\sum_{i=1}^{m} P_i}{m} \quad (4)$$

$$\hat{R}^M = \frac{\sum_{i=1}^{m} R_i}{m} \quad (5)$$

$$\hat{F}^M = \frac{2\left(\hat{P}^M\right)\left(\hat{R}^M\right)}{\hat{P}^M + \hat{R}^M} \quad (6)$$

Table 2 and Fig. 4 illustrate the results for classification accuracy. From the results, we compare the effectiveness of employing the ontology enrichment algorithm for classifying biomedical text abstracts. Generally, the experimental results indicate that the hierarchical classification method with the ontology enrichment algorithm outperforms the hierarchical classification method without the ontology enrichment algorithm.

In addition, the hierarchical classification experiments with ontology enrichment using 2,132 features that combined the MeSH entry terms and 12 heart disease glossaries achieve top-notch performance, with the macro-average for precision, recall and F measure being 100%, 56% and 71.8%, respectively. It is clear that the ontology enrichment process using the features from combination of MeSH entry terms and heart disease glossaries performs significantly better than other biomedical information sources. This might be because most of the selected features in this experiment are relevant and important features for enriching the initial training ontology.

Table 2: Results on hierarchical classification using different biomedical information sources

Experiments	Macro-Average		
	Prec.	Recall	F
Without Enrichment	74.1	36.2	48.6
OHSUMED	91.4	43.3	58.8
MeSH (Baseline)	82.7	38.4	52.4
Glossary	100	53.0	69.3
MeSH + Glossary	100	56.0	71.8

On the other hand, we observe that the ontology enrichment using 892 MeSH entry terms produces the worst performance compared to other biomedical information sources, producing a macro-average of 82.7% for precision, 38.4% for recall and 52.4% for F measure, respectively. The low performance of this experiment might be because some features that are selected and collected from the MeSH tree structures unmatched the features in the testing documents. This is due to the limitation of terms or feature variants and some synonym features that are extracted and used for enriching the training ontology.

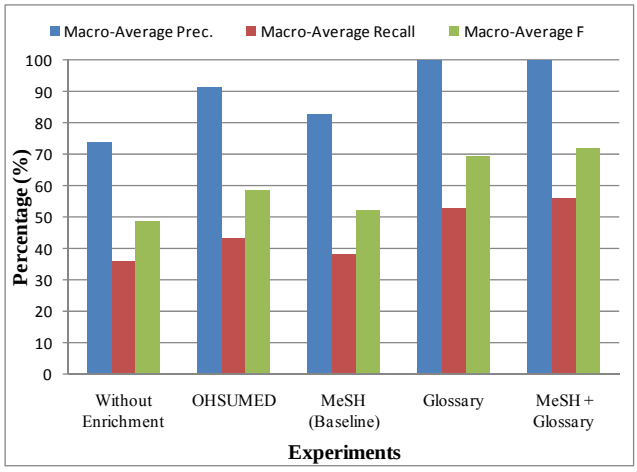

Fig. 4 Performance of hierarchical classification with and without employing the ontology enrichment algorithm.

In addition, although the result on the 7,565 features from the OHSUMED is not poor, the response speed of the algorithm is slow, while the result on the glossary experiments can be regarded as another example of good performance. We find that the macro-average for precision, recall and F measure obtained by the hierarchical classification experiments using 1,362 features from 12 heart disease glossaries are a little higher than those obtained by the hierarchical classification experiments using 7,565 features from the OHSUMED dataset. This indicates that the number of features used for expanding the training ontology does not greatly influence the performance of the hierarchical classification method. In addition, the selected features from the OHSUMED dataset used for enriching the ODO may contain general features, which lead to misclassification.

6. Discussion and Conclusion

In this paper, we present and evaluate the effectiveness of our proposed ontology enrichment algorithm for classifying biomedical text abstracts. Consequently, we conduct several experiments using the features that are extracted from different biomedical information sources such as a collection of biomedical text abstracts (from the OHSUMED dataset), MeSH entry terms and heart disease glossaries. Generally, the experimental results indicate that our proposed ontology enrichment algorithm can predict more specific categories for classifying biomedical text abstracts which improve the classification performance.

Overall, the results show the different performance on the hierarchical classification experiments. We find that the performance of hierarchical classification with the ontology enrichment algorithm outperforms the performance of hierarchical classification without the ontology enrichment algorithm. Furthermore, we observe that the ontology enrichment algorithm using features combined from MeSH entry terms and heart disease glossaries achieves the best performance in hierarchical classification experiments compared to other biomedical information sources. In addition, the results that are obtained in the experiments indicate that the performance of hierarchical classification method does not respond to the number of features used for expanding the initial training ontology. For our future target, we intend to investigate and explore the effectiveness of the instance matching approach for ontology enrichment which might achieve higher recall in hierarchical classification experiments.

References

[1] PubMed, http://www.ncbi.nlm.nih.gov/pubmed/ (accessed May 2014).
[2] K. Thaoroijam, "A Study on Document Classification using Machine Learning Techniques", International Journal of Computer Science Issues, Vol. 11, No. 1, 2014, pp.217-222.
[3] G. Antoniou, and F. V. Harmelen, "A Semantic Web Primer", London: The MIT Press, 2004.
[4] M. K. Geeta, "Automatic Hierarchical Classification of Documents", Master's thesis, India: Indian Institute of Technology, 2001.
[5] A. Pulijala, and S. Gauch, "Hierarchical Text Classification", in Proceedings of the International Conference on Cybernetics and Information Technologies (CITSA), 2004, pp. 21-25.

[6] S. Gauch, A, Chandramouli, and S. Ranganathan, "Training a Hierarchical Classifier using Inter-Document Relationships", Technical Report, ITTC-FY2007-TR-31020-01, 2006.
[7] T. Li, S. Zhu, and M. Ogihara, "Hierarchical Document Classification using Automatically Generated hierarchy", Journal of Intelligent Information Systems, Vol. 29, No. 2, 2007, pp.211-230.
[8] 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, 2006, pp. 33-40.
[9] 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 on Research and Development in Information Retrieval, 2008, pp. 619-626.
[10] M. Warin, H. Oxhammar, and M. Volk,"Enriching an Ontology with WordNet based on Similarity Measures", in MEANING-2005 Workshop, 2005.
[11] M. Speretta, and S. Gauch, "Using Text Mining to Enrich the Vocabulary of Domain Ontologies", in ACM International Conference on Web Intelligence, 2008, pp.549-552.
[12] P. Arnold, and E. Rahm, "Semantic Enrichment of Ontology Mappings: A Linguistic-based Approach", Advances in Databases and Information Systems, Lecture Notes in Computer Science, Vol. 8133, 2013, pp. 42-55.
[13] R. B. Dollah, and M. Aono, "Classifying Biomedical Text Abstracts based on Hierarchical 'Concept' Structure", in International Conference on Data Mining and Knowledge Engineering, 2011, pp.513-518.
[14] R. B. Dollah, and M. Aono, "Ontology based Approach for Classifying Biomedical Text Abstracts", International Journal of Data Engineering (IJDE), Vol.2, Issue 1, 2011, pp.1-15.
[15] H. P. Luong, S. Gauch, and M. Speretta, "Enriching Concept Descriptions in an Amphibian Ontology with Vocabulary Extracted from WordNet", in The 22nd IEEE International Symposium on Computer-Based Medical Systems, 2009, pp.1-6.
[16] A. Sun, and E.P. Lim, "Hierarchical Text Classification and Evaluation", in Proceeding of the 2001 IEEE International Conference on Data Mining, 2001, pp. 521-528.
[17] W. Hersch, C. Buckley, T. J. Leone, and D. Hickam, "OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research", in Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994, pp.192-201.
[18] 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, Vol. 7, pp. 344-356.
[19] MeSH Tree Structures, http://www.nlm.nih.gov/mesh/trees.html (accessed January 2013).
[20] Heart Health Center. Glossary of Terms Used in Heart Disease and Cardiology, http://heartdisease.about.com/cs/glossary/a/glossary.htm (accessed April 2013).
[21]Glossary of Heart Disease Terms, http://newsok.com/glossary-of-heart-disease-terms/article/3405169 (accessed April 2013).
[22] Heart Health Center. Heart Disease Glossary of Terms, http://www.webmd.com/heart-disease/heart-disease-glossary (accessed April 2013).
[23] Emedicine Health. Heart Disease Glossary of Medical Terms,http://www.emedicinehealth.com/coronary_heart_disease/ glossary_em.htm (accessed April 2013).
[24] EhealthMD, "Glossary", http://ehealthmd.com/library/heartdisease/HD_glossary.html (accessed April 2013).
[25] Effient, "Glossary", http://www.effient.com/Pages/heart-disease-glossary.aspx (accessed April 2013).
[26] British Heart Foundation, "Glossary-Heart terms", http://www.yheart.net/default.aspx?page=126 (accessed April 2013).
[27] California Pacific Medical Center, "Glossary and Definitions",http://www.cpmc.org/advanced/heart/patients/topics/eglossary.html (accessed April 2013).
[28] Health Dictionary, "Heart Terms Listed Alphabetically", http://www.health-dictionary.com/heart (accessed April 2013).
[29] Cleveland Clinic,"Dictionary", http://my.clevelandclinic.org/heart/glossary/a.aspx. (accessed April 2013).
[30]Glossary Terms for Heart Disease. http://www.cardiology.net/glossary/Glossary-Terms-for-Heart-Disease (accessed April 2013).
[31] Providence Health & Services, "Glossary of Heart and Vascular Terms", http://www.providence.org/everett/health_resource_centers/heart_disease_center/glossary.htm. (accessed April 2013).

Rozilawati binti Dollah received her B.Sc. and M.Sc. degrees from Universiti Teknologi Malaysia, both in Computer Science in 1998 and 2001. She is a lecturer in the Department of Information Systems, Universiti Teknologi Malaysia. She is currently a Ph.D. candidate at the Graduate School of Electronic and Information Engineering, in Toyohashi University of Technology. Her research interests are data mining, text mining and semantic web.

Masaki Aono received his B.Sc. and M.Sc. degrees from University of Tokyo in 1981 and 1984. He completed his Ph.D. from Rensselaer Polytechnic Institute, New York in 1994. He has joined the IBM Research, Tokyo Research Laboratory from 1984 until 2003. Since 2003, he is a Professor of Information and Computer Sciences Department, in Toyohashi University of Technology. His research interest are massive multimedia datasets, data mining, web mining, semantic web include feature extraction, classification, clustering, segmentation and information retrieval. He is a member of ACM, IEEE Computer Society, IPSJ, IEICE, JSAI and NLP.