International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:2, 2011

Classifying Biomedical Text Abstracts based on Hierarchical 'Concept' Structure

Rozilawati Binti Dollah and Masaki Aono

***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. In this paper, we present an approach for classifying a collection of biomedical text abstracts downloaded from Medline database with the help of ontology alignment. To accomplish our goal, we construct two types of hierarchies, the OHSUMED disease hierarchy and the Medline abstract disease hierarchies from the OHSUMED dataset and the Medline abstracts, respectively. Then, we enrich the OHSUMED disease hierarchy before adapting it to ontology alignment process for finding probable concepts or categories. Subsequently, we compute the cosine similarity between the vector in probable concepts (in the "enriched" OHSUMED disease hierarchy) and the vector in Medline abstract disease hierarchies. Finally, we assign category to the new Medline abstracts based on the similarity score. The results obtained from the experiments show the performance of our proposed approach for hierarchical classification is slightly better than the performance of the multi-class flat classification.

***Keywords*—**Biomedical literature, hierarchical text classification, ontology alignment, text mining.

I. 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 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 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.

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 [2] stated that the flat classification approach was suitable when a small number of categories were defined. However, 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 much preferred.

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. Recently, the use of hierarchies for text classification has been widely investigated and applied by many researchers. However, only little attention has been paid to the classification of biomedical literature. Therefore, in this paper, we propose a hierarchical classification method where we employ the hierarchical 'concept' structure for classifying biomedical text abstracts with the help of ontology alignment.

Our proposed method is different compared to the previous works because we construct two types of hierarchies, which are the "enriched" OHSUMED disease hierarchy as our training ontology and the Medline abstract disease hierarchy as testing ontology. Next, we employ the ontology alignment process to match the concepts and relations between the "enriched" OHSUMED disease hierarchy and the Medline abstract disease hierarchy for exploring and searching the aligned pairs. We select 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 relevant category. Finally, we classify this abstract based on the similarity score.

In order to evaluate our approach, we conducted the experiments of the multi-class flat classification as our baseline.

Then, we compare the results of our proposed method with the 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 II, we summarize the related work on text classification. The proposed hierarchical classification method is explained in section III. Section IV contains experiments and results. Finally, we conclude this paper with a summary and suggestions for future work in section V.

II. Related Work

Text classification is the process of using automated techniques to assign text samples into one or more set of predefined classes [3]. In flat classification, the classifier will assign a new documents to a category based on training

R. B. Dollah is a PHD student in the Graduate School of Engineering, Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan and is with the Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Johor, Malaysia (e-mail: rozeela@kde.cs.tut.ac.jp, rozilawati@utm.my.

M. Aono is a professor in the Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan (e-mail: aono@tut.ac.jp).

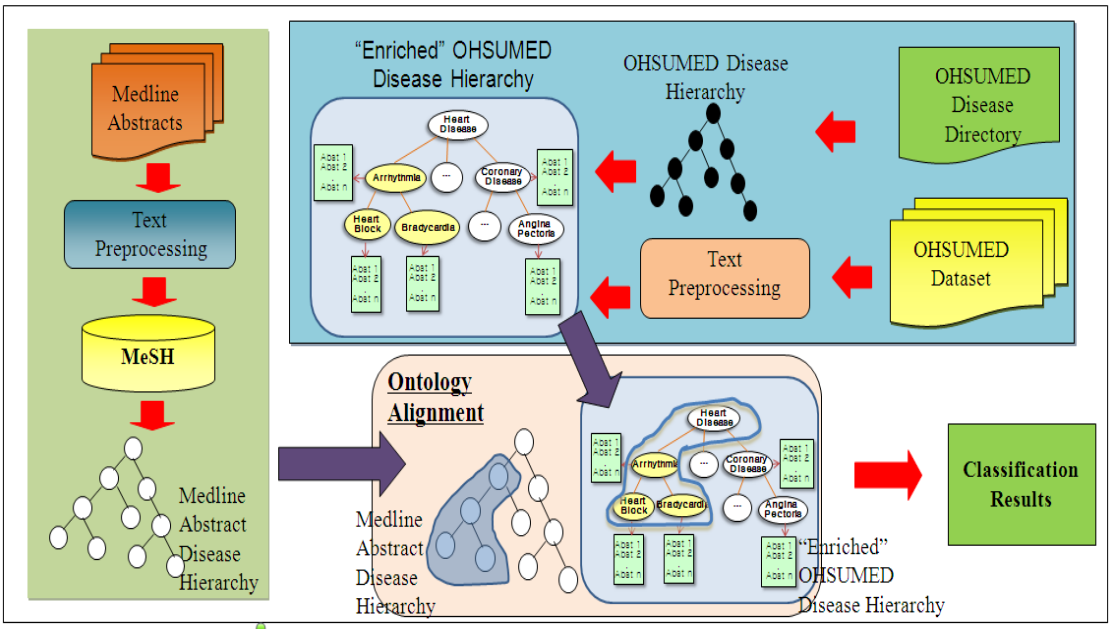

Fig. 1 A method for hierarchical classification of biomedical text abstracts

examples of predefined documents. Meanwhile, 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.

A few hierarchical classification methods have been proposed. For example, Pulijala and Gauch [4] and Gauch, et al. [5] reported that they classify documents during indexing which can be retrieved by using a combination of keyword and conceptual match. Ruiz and Srinivasan [6] have proposed a text categorization method based on the Hierarchical Mixture of Expert (HME) model using neural networks. Li et al. [7] have proposed another approach of hierarchical document classification using linear discriminant projection to generate topic hierarchies.

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 [6] and support vector machine (SVM) [8], [9], [10]. SVM has been prominently and widely used for different classification task, in particular for document classification. For instance, Nenadic et al. [8] 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 were used the voting score from all category-to-category classifier for assigning a web document to a sub-category. And they reported that their method can improve the performance of imbalanced data.

In [10], Dumais and Chen reported that they used the hierarchical structures for classifying web content. Accordingly, they were employed SVM to train second-level category models using different contrast sets. Eventually, they classified a web content based on the scores that were combined from the top-level and second-level model.

Our approach is different from the previous works, where we explore the use of hierarchical 'concept' structure with the help of ontology alignment for searching and identifying the probable categories in order to classify biomedical text abstracts.

III. Hierarchical Classification Method

The large number of biomedical articles that published on web has made the process of classification becomes challenging and difficult. Lately, various classification methods are proposed for classifying biomedical literature.

However, in our research, we explore the use of hierarchical structure or ontology for classifying biomedical text abstracts. 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. Fig. 1 illustrates our proposed hierarchical classification method that implemented in our research.

A. Datasets

For our experiments, we use two different datasets, which are a subset of the OHSUMED dataset as training documents and Medline abstracts as test documents. The OHSUMED dataset [11] is a subset of clinical paper abstracts from the Medline database, from year 1987 through 1991. We select 400 records (documents) from 43 different categories from the subset of OHSUMED dataset for enriching the OHSUMED disease hierarchy.

While, for classification purpose, we have selected randomly a subset of Medline abstracts from the Medline database. We retrieved this dataset with the query terms, such as "arrhythmia", "heart block", etc. Then, we indexed this dataset belonging to 12 categories of disease as shown in Table 1. A total number of Medline abstracts are 100.

TABLE 1
THE NUMBER AND CATEGORIES OF MEDLINE ABSTRACTS

Category No.	*Category Name*	*No. of Documents*
1	Arrhythmia	10
2	Heart Block	5
3	Coronary Disease	10
4	Angina Pectoris	5
5	Heart Neoplasms	10
6	Heart Valve Diseases	10
7	Aortic Valve Stenosis	5
8	Myocardial Diseases	10
9	Myocarditis	10
10	Pericarditis	5
11	Cancer	10
12	Human Disease	10
	Total	**100**

B. Text Preprocessing and Feature Selection

In our research, we perform text preprocessing for extracting the features (or noun phrases) from the OHSUMED dataset and the Medline abstracts, respectively, 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. Meanwhile, phrase chunking is the process of recovering the noun phrases constructed by the part-of-speech tags. Then, we employed these noun phrases (as concepts) for constructing the Medline abstract disease hierarchy and enriching the OHSUMED disease hierarchy. In addition, we also perform feature selection in order to reduce the original features to a small number of features by removing the rare and irrelevant features.

In feature selection process, we employ the document frequency and the chi-square (χ^2) techniques to distinguish between relevant and irrelevant features. Firstly, 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. Next, we eliminate the features with the highest and lowest frequencies. The purpose of this process is to 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. 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.

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

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

For 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. For our experiments, we only select the features with the χ^2 score greater than 3.841 as the relevant features. We use all of selected features for enriching the OHSUMED disease hierarchy and also for constructing the Medline abstract disease hierarchies. Table 2 below shows the statistic of the relevant features that produced in the feature selection process.

TABLE II
THE STATISTIC OF SELECTED FEATURES

	Number of Features
Features (after extract noun phrases)	3,145
After remove DF<2 and DF>20	1,130
After remove Chi-square score < 3.841	1,081

Each document is represented by the TF-IDF weighting. Therefore, we calculate each term or feature weight using the equation given below.

$$w_{i,d} = tf_{i,d} \log\left(\frac{n}{df_i}\right) \quad ; \quad \frac{n+1}{df_i+1} \qquad (2)$$

where term frequency $tf_{i,d}$ is the frequency of term i occurs in document j and $d = 1, \ldots, m$. Document frequency df_i is the total number of documents that contain term i and n is the total number of documents.

C. "Enriched" OHSUMED Disease Hierarchy

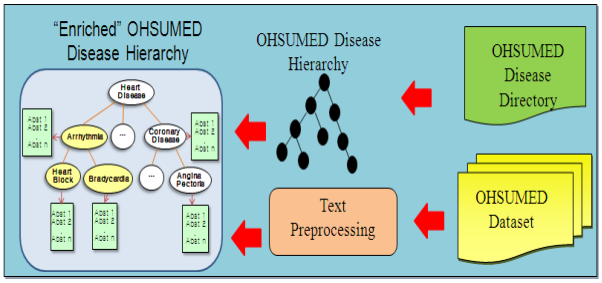

Fig. 2 An approach for constructing and enriching the OHSUMED disease hierarchy

In our research, we construct the OHSUMED disease hierarchy by referring to the OHSUMED disease directory. Subsequently, we enrich this hierarchy by assigning the features that extracted from the OHSUMED dataset to each node of the hierarchy as shown in Fig. 2.

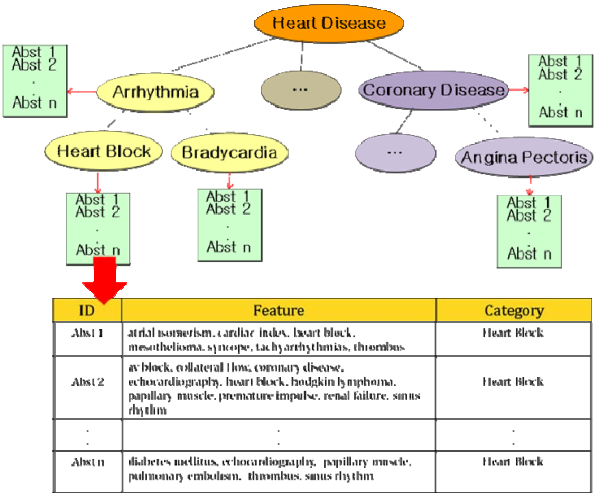

Fig. 3 An example of the "enriched" OHSUMED disease hierarchy

The important task in enriching the OHSUMED disease hierarchy is to select meaningful and relevant features from the

OHSUMED dataset. Enrichment had been done by mapping each concept or node in the OHSUMED disease hierarchy with a set of related features that extracted and selected from the OHSUMED dataset. Fig. 3 describes the example of the "enriched" OHSUMED disease hierarchy.

D. Medline Abstract Disease Hierarchy

The Medline abstract disease hierarchies were constructed using the selected features that extracted from a collection of biomedical text abstracts that downloaded from the Medline database. The description of text preprocessing and feature selection has been explained in *subsection B*.

In our proposed approach, we create the Medline abstract disease hierarchy using Protégé. For this purpose, we refer to the Medical Subject Headings (MeSH) for indexing our features. We map all selected features to the concepts in the MeSH tree structure [12] for identifying heading and subheading of hierarchical grouping. Finally, we construct the Medline abstract disease hierarchy by using the heading and subheading of hierarchical grouping that are suggested by the MeSH tree structure. Fig. 4 depicts a part of the Medline abstract disease hierarchy.

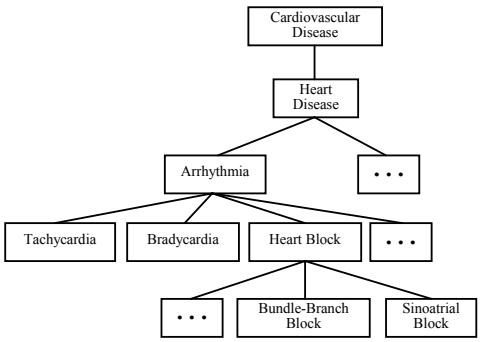

Fig. 4 An example of a part of the Medline abstract disease hierarchy

E. Ontology Alignment

The purpose of ontology alignment in our research is to match the concepts and relations between the "enriched" OHSUMED disease hierarchy and the Medline abstract disease hierarchy. For that reason, we perform ontology alignment using the "Anchor-Flood" algorithm (AFA) [13]. 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 "enriched' OHSUMED disease hierarchy and the Medline abstract disease hierarchy for producing the aligned pairs (aligned entities across hierarchies or ontologies). 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 select all of the aligned pairs as our probable concepts for classification purpose.

F. Ontology-based Hierarchical Classification

In our proposed approach, we construct two types of hierarchies, which are the OHSUMED disease hierarchy (as our training ontology) and the Medline abstract disease hierarchies (as testing ontology). Then, we perform ontology alignment in order to match both hierarchies for producing the aligned pairs. We consider the aligned pairs as a set of probable relevant categories for classifying biomedical text abstracts.

Afterwards, we evaluate the more specific concepts based on the similarity between the new Medline abstract and probable relevant category. We compute the cosine similarity score between the vector of unknown new abstract in each Medline abstract disease hierarchy and the vector of each probable category in the "enriched" OHSUMED disease hierarchy for identifying and predicting more specific category. The cosine similarity score would be calculated using the following equation.

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

where the vector of $d_j = (w_{11}, w_{12}, \ldots, w_{1n})$ and the vector of $d_k = (w_{21}, w_{22}, \ldots, w_{2n})$. We rank all the probable categories based on their cosine similarity score. Eventually, we classify the new Medline abstracts into the first rank of similarity score.

IV. EXPERIMENTS AND RESULTS

For our experiments, we use 400 records (documents) from 43 different categories of the subset of OHSUMED dataset for enriching our ontology learning. While, for classification purpose, we randomly downloaded 100 biomedical text abstracts that related to human diseases from Medline database. Then, we conducted the multi-class flat classification experiments using LIBSVM [14], which ignore hierarchical structure as our baseline for evaluating the performance of our proposed method for hierarchical classification of biomedical text abstracts. The results and the performance of the flat and hierarchical classification are shown in Table 3 and Fig. 5.

TABLE III
THE RESULTS OF THE FLAT AND HIERARCHICAL CLASSIFICATION

Category No.	Flat Classification (% Accuracy)	Flat Classification with chi-square (% Accuracy)	Hierarchical Classification (% Accuracy)	Hierarchical Classification with chi-square (% Accuracy)
1	0	20	50	50
2	0	0	0	0
3	0	0	0	0
4	0	10	0	0
5	20	50	30	20
6	10	10	0	0
7	0	0	10	10
8	0	20	40	50
9	10	20	0	0
10	10	0	30	40
11	0	0	0	0
12	0	0	0	0
Average (% Accuracy)	5	13	16	17

Confusion matrix in Table 4 demonstrates the details of the results of the hierarchical classification, where we used the features that are selected from the feature selection process.

TABLE IV
CONFUSION MATRIX FOR THE HIERARCHICAL CLASSIFICATION (WITH CHI-SQUARE)

ACTUAL CLASS \ PREDICTED CATEGORY	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	UNKNOWN	Total
C1	**5**	0	0	0	0	0	0	1	1	0	0	0	3	10
C2	0	**0**	0	0	0	0	0	0	0	0	0	0	5	5
C3	0	0	**0**	0	0	0	0	0	0	0	0	0	10	10
C4	0	0	0	**0**	0	0	0	0	0	0	0	0	5	5
C5	0	0	0	0	**2**	0	0	0	0	0	0	0	8	10
C6	0	0	0	0	0	**0**	0	0	0	0	0	0	10	10
C7	0	0	0	0	0	0	**1**	0	0	0	0	0	4	5
C8	0	0	0	0	0	0	0	**5**	0	0	0	0	5	10
C9	0	0	0	0	0	0	0	0	**0**	0	0	0	10	10
C10	0	0	0	0	0	0	0	0	0	**4**	0	0	1	5
C11	0	0	0	0	0	0	0	0	0	0	**0**	0	10	10
C12	0	0	0	0	0	0	0	0	0	0	0	**0**	10	10
UNKNOWN	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	5	0	0	0	2	0	1	6	1	4	0	0	81	100

17%	17 of 100 correct
83%	83 of 100 incorrect or unknown

* C is Category

Overall, our proposed approach performs slightly better that the baseline. We found the experimental results indicate that our proposed approach performs slightly better than the flat classifier using SVM, whereby the performance of the flat and hierarchical classification show on average 13% and 17% accuracy, respectively. Based on the results that obtained from the experiments, overall, we found the classification accuracies in some of categories in the hierarchical classification experiments, such as in category 1 (arrhythmia), 7 (bundle-branch block), 8 (coronary thrombosis) and 10 (aortic coarctation) perform better than in the flat classification experiments.

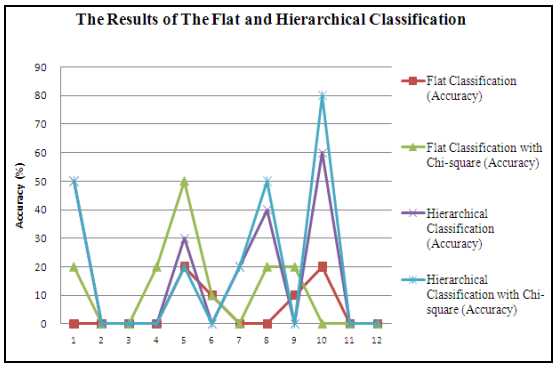

Fig. 5 The performance of classification accuracies

In addition, the classification accuracies of the category 1 (arrhythmia), 5 (heart valve diseases), 8 (coronary thrombosis) and 10 (aortic coarctation) perform better in both flat and hierarchical classification experiments compared to other diseases categories. This might be caused by some of the Medline abstracts too short and would extract very few relevant features. Consequently, we could construct a small hierarchy for ontology alignment purpose, which may produce a small number of aligned pairs (probable category).

Besides that, the small number of documents that are represented in a particular class in the dataset may also affect the decrease of the classification accuracy. Even though the percentage of performance improvement in the hierarchical classification is small compared to the performance of the flat classification, we believe that our proposed approach suggests that by using hierarchical 'concept' structure with the help of ontology alignment could improve the classification performance.

V. CONCLUSION

In this paper, we proposed a hierarchical classification method that utilized the ontology alignment, namely 'Anchor-Flood' algorithm to search the probable categories for given biomedical text abstracts. Then, we evaluated the performance of our approach by conducting the hierarchical classification experiments using the features that extracted from the OHSUMED dataset and Medline abstracts. As our baseline, we performed the multi-class flat classification experiments using SVM. Other than that, we also performed feature selection by employing the document frequency and

chi-square techniques for selecting relevant features. Then, we conducted some experiments of the flat and hierarchical classification using the features that are selected from feature selection process.

Generally, the experimental results indicate that the use of hierarchical 'concept' structure with the help of 'Anchor-Flood' algorithm can improve 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 concept 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 and expand the OHSUMED disease hierarchy and Medline abstract disease hierarchies. By increasing the total number of documents that are represented in each class in the dataset may lead to performance increase of our proposed approach for hierarchical text classification.

ACKNOWLEDGMENT

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.

REFERENCES

[1] F. M. Couto, B. Martins and M. J. Silva, "Classifying biological articles using web sources", In Proceedings of the 2004 ACM symposium on Applied Computing, 2004, pp. 111-115.
[2] 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.
[3] A. M. Cohen, "An effective general purpose approach for automated biomedical document classification", AMIA 2006 Symposium Proceeding, 2006, pp. 161-162.
[4] A. K. Pulijala and S. Gauch, "Hierarchical text classification", 2004, URL: http://academic.research.microsoft.com/Paper/12788733.aspx.
[5] S. Gauch, A. Chandramouli and S. Ranganathan, "Training a hierarchical classifier using inter-document relationships", Technical Report, ITTC-FY2007-TR-31020-01, August 2006.
[6] M. E. Ruiz and P. Srinivasan, "Hierarchical neural networks for text categorization", Information Retrieval, 5, 2002, pp. 87-118.
[7] T. Li, S. Zhu and M. Ogihara, "Hierarchical document classification using automatically generated hierarchy", Journal of Intelligent Information Systems, 29(2), 2007, pp. 211-230.
[8] 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, Vol. 13, 2003, pp. 121-128.
[9] Y. Wang and Z. Gong, "Hierarchical classification of web pages using support vector machine", In Proceedings of 11th International Conference on Asian Digital Libraries, ICADL 2008, Bali, Indonesia. Proceedings, Lecture Notes in Computer Science 5362, Springer, 2008, pp. 12-21.
[10] S. Dumais and H. Chen, "Hierarchical classification of web content", In Proceeding of the SIGIR2000, Athens, GR, 2000, pp. 256-263.
[11] OHSUMED dataset, URL: http://davis.wpi.edu/xmdv/datasets/ohsumed.html.
[12] Medical Subject Heading (MeSH) tree structures, URL: http://www.nlm.nih.gov/mesh/trees.html.
[13] 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), 2009, pp. 344-356.
[14] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines", 2007, URL: http://www.csie.ntu.edu.tw/~cjlin/libsvm.