Two-Step Text Classification for Multi-Label Drug Monographs

Verayuth Lertnattee and Chanisara Lueviphan

***Abstract*—With an increasing number of documents for drug information, automatic classification of documents is an important task for organizing these documents into appropriated classes. Only a few research works on text classification of drug documents, were contributed. In this task, monographs of drug can be categorized to one class or multiple classes by their indications. A two-step text classification for multi-label documents is proposed. A set of drug monographs is drawn from RxList and DailyMed websites and used for evaluating the proposed method. In the first step, documents are classified as single-label or multi-label documents. The result from the first step is applied on the second step of classification based on their therapeutic classes. From the experimental results, the two-step classification is an efficient method to classify a document into one or multiple classes.**

***Index Terms*—Text categorization, text classification, multi-label classification, drug monograph classification.**

I. Introduction

Nowadays, the online information is increasing rapidly including drug information. The health care teams and the patients, access to the online drug information and use it frequently, in order to improve therapeutic efficiency of drug selection and use. As a common type of drug information, drug monographs are widely used to convey details of drugs. The topics in drug monograph generally include brand names, chemical names, generic names, descriptions, clinical pharmacology, indication, and so on. The sparse online drug information is a type of drug monograph and defines therapeutic class [1]. A pharmacist could really do with therapeutic class for treatment, diseases, and symptoms. Thus classification techniques in therapeutic class of online drug monograph need pharmacist to determine the best and most efficient search strategy. Due to huge of drug information used by health care professionals and patients, it is hard to classified or organized groups of drug documents by professionals. Automatic text categorization (or text classification) is an important tool in efficiently organizing text document. Given a training set of label documents, text classification is a supervised learning method to use information from the training set to assign a class (or classes) to a document. For a drug monograph, it is often assigned only one class. However, for some monographs, they are assigned more than one class., i.e., some drugs may have more than one of therapeutic groups. In this paper, a technique called two-step classification is proposed to classify a set of drug monograph which a monograph is either one category or multiple categories according to its content. Performance of the proposed method is investigated on drug monographs, collected from two websites, i.e., RxList and DailyMed. This paper is organized as follows. Section 2 shows detail in drug monograph. Section 3 presents multi-label text classification and centroid-based classifier. Section 4 is experimental setting and evaluation. Two-step text classification is described in Section 5. In section 6, experimental results were reported. Finally, Section 7 is the conclusion and future work.

Manuscript received January 9, 2013; revised March 12, 2013. This work was supported in part by the Research and Development Institute, Silpakorn University via research grant SURDI 53/01/12.

The authors are with the Faculty of Pharmacy, Silpakorn University, Nakorn Pathom, 73000 Thailand (e-mail: verayuth@su.ac.th, verayuths@hotmail.com, chanisara@su.ac.th).

II. Drug Monograph

Drug information is consists of drug data that has indication, administration, adverse drug reaction, and etc. A common type of drug information that collects drug description, called drug monographs are widely used by health care professionals and patients. Furthermore, drug monographs are often arranged into therapeutic classes or generic names of the drugs. The topics in monographs generally include brand names, chemical name, generic names, descriptions, clinical pharmacology, indications, dosage, administration, interaction, contraindications, adverse effects, over dosage, and so on. The main resources of drug monographs come from the standard references, e.g., American Hospital Formulary Service Drug Information (AHFS Drug information), Facts and Comparisons, Physician's Desk Reference, and Mosby's Drug Consult. Nowadays, websites provide health information is expanding broadly. AHFS Drug information is one of the most popular textbook for pharmacist. Therapeutic classification in AHFS Drug Information is a cause of popularity because pharmacists solve the problem quickly. Some of website providers need to service the health care teams and the patients, and then online drug information is developed. Some well known websites for providing drug monographs such as RxList, DailyMed, MedicineNet.com and etc., are commonly used in practice. To increase the efficiency and the accurate information we search, the online drug information providers need to classify the web pages of these websites into therapeutic classes. Automatic text categorization is a valuable tool for organizing documents into a class or classes based on their content. It can be applied to a website or a search engine for drug information. It can be used for filter a set of desirable documents. With the best of our knowledge, there is no research work contribute to multi-label text classification for drug information.

DOI: 10.7763/IJIMT.2013.V4.405

III. Multi-label Text Classification and Centroid Based Classifier

With the increasing availability of online information, text classification turns into the important techniques by using machine learning. The objective of machine learning is to learn classifiers from example which perform the category assignments automatically. This type of learning is induction-based supervised concept learning or just supervised learning. The supervised learning is a part of data mining, the process of employing one or more computer learning techniques to automatically analyze and extract knowledge from data contained within a database [2]. Therefore, text classification falls within the machine learning paradigm and data mining. The definition of text classification is the activity of labeling natural language texts with thematic categories from a predefined set [3]. Several researches on text classification contributed to single-label classification. However, this paper focuses on multi-label text classification. Classification techniques have been developed in a variety of learning techniques such as probabilistic models [4], neural network [5], example-based models (e.g., k -nearest neighbor) [6], linear models [7], [8], support vector machine [9] and so on. Among these methods, a linear model called a centroid-based approach is attractive since it has relatively less computation than other methods in both the learning and classification stages. Despite less computation time, centroid-based methods were shown in several literatures including those in [7], [8], to achieve relatively high classification accuracy. In a centroid-based method, an individual class is modeled by weighting terms appearing in training documents assigned to the class. This makes classification performance strongly depend on term weighting applied in the model. Most previous works of centroid-based classification focused on weighting factors related to frequency patterns of terms or documents in the class. For the rest of this section, details of multi-label text classification and the centroid-based classifier are given.

A. Multi-Label Text Classification

For a single-label text classification, a document is assigned only one category. Two approaches of classification are utilized to handle single-label classification, i.e., binary classification or multi-class classification. However, the problems in real work usually fall into the problem of multi-label text categorization, where each text document is assigned to one or more categories. Existing methods for multi-label classification can be divided into two main methods, i.e., problem transformation methods and algorithm adaptation methods [10]. Problem transformation methods can be defined as methods that transform the multi-label classification problem either into one or more single-label classification problems or regression problems. This is the same solution to solve the problem of single-label classification using a binary classifier. However, the binary classification using a binary classifier based on the assumption of label independence. Therefore, during its transformation process, this method ignores label correlations that exist in the training data. Due to this information loss, predicted label sets from the binary classification are likely to contain either too few or too many labels, or labels that would never co-occur in practice [11]. With some limitations of binary classification, some extensions had been done such as [11, 12] to provide better performance of classification.

For algorithm adaptation methods, they can be defined as methods that extend or modify specific learning algorithms in order to handle multi-label data directly. The examples for these methods are shown in [13, 14]

B. Centroid-Based Classifier

In the centroid-based text categorization, a document (or a class) is represented by a vector using a vector space model with a bag of words (BOW) [15]. The simplest and popular one is applied term frequency (*tf*) and inverse document frequency (*idf*). It is usually used in the form of $tf \times idf$ as a term weight for representing a document. In a vector space model, given a set of documents $D = \{d_1, d_2, ..., d_{|D|}\}$, a document d_j is represented by a document vector $\vec{d}_j = \{w_{1j}, w_{2j}, ..., w_{|T|j}\} = \{tf_{1j} \times idf_1, tf_{2j} \times idf_2, ..., tf_{|T|j} \times idf_{|T|}\}$, where w_{ij} is a weight assigned to a term t_i in a set of terms (T) of the document. In this definition, tf_{ij} is term frequency of a term t_i in a document d_j and idf_i is inverse document frequency, defined as $log(|D|/df_i)$. Here, $|D|$ is the total number of documents in a collection and df_i is the number of documents, which contain the term t_i . Besides term weighting, normalization is another important factor to represent a document or a class. Class prototype $\vec{c}_k$ is obtained by summing up all document vectors in C_k and then normalizing the result by its size. The formal description of a class prototype $\vec{c}_k$ is $\sum_{dj \in C_k} \vec{d}_j / \| \sum_{dj \in C_k} \vec{d}_j \|$, where $C_k = \{ d_j | d_j$ is a document belonging to the class $c_k \}$. The simple term weighting is $\overline{tf} \times idf$ where $\overline{tf}$ is an average class term frequency of the term. The formal description of $\overline{tf}$ is $\sum_{dj \in C_k} tf_{ijk} / |C_k|$, where $|C_k|$ is the number of documents in a class c_k . Term weighting described above can also be applied to a query or a test document. In general, the term weighting for a query is $tf \times idf$. Once a class prototype vector and a query vector have been constructed, the similarity between these two vectors can be calculated. The most popular one is cosine distance [16]. This similarity can be calculated by the dot product between these two vectors. Therefore, the test document will be assigned to the class whose class prototype vector is the most similar to the vector of the test document. In this paper, a multi-label document is also taken into account. A ranking categorization is applied for this work. Given a document d_j, a system ranks the categories in $C = \{c_1, c_2, ..., c_{|C|}\}$ according to their estimated similarity to document d_j. A ranked list of possible categories will be considered.

Some previous works, such as those in [8], attempted to apply some factors called term distribution factors to improve performance of a centroid-based classifier with the basic term weighting of $tf \times idf$. In this paper, a centroid-based classifier with a term distribution factor., i.e., standard deviation of a term (*sd*), is used along with the standard *tfxidf*.

IV. Two-Step Text Classification

Our method is an algorithm adaption method. In this approach of classification, two steps of classification are divided. The first step is to assign the correct numbers of classes to a document. The last step is to assign a real world class (or classes) to a document based on the list of score generated by a classifier. The algorithm of two-step classification is shown in Fig. 1 and decribed as follow.

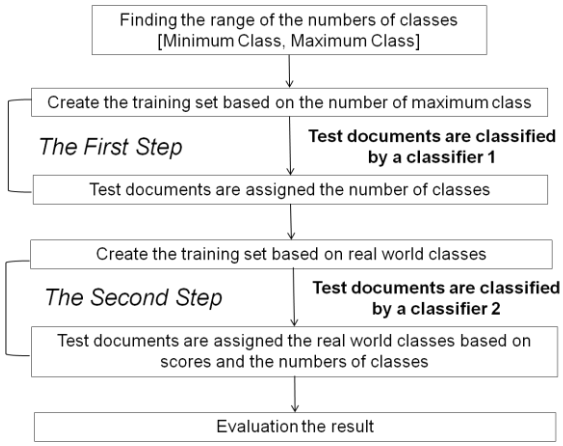

Fig. 1. The algorithm of two-step multi-label classification.

1) The data set needs to have a finite number of real world classes. In drug monographs, according to AHFS Drug Information 2012, a drug monograph has at least one therapeutic class. The maximum number of classes is three. Therefore, the range of the numbers of classes is [1], [3]
2) The training set is created based on the maximum number of classes, i.e, 3 for drug monographs. A classifier 1 is select to train from the training set. In this work, a centroid-based classifier is used.
3) A set of test documents is assigned the numbers of real world classes using classifier 1.
4) The training set is created based on the real world classes. In this work, 11 therapeutic classes of drug monographs is applied. A classifier 2 is select to train from the training set. In this work, a centroid-based classifier is used, the same as classifier 1.
5) The same set of test documents is assigned more or more of real world classes. A set of top highest score values from sorting list of possible classes, are selected based on the number of classes for that documents on the first step.
6) Evaluation the result based on evaluation criteria.

V. Experimental Settings and Evaluation

In this section, 2 topics are more detail described, i.e., experimental setting for drug monographs and evaluation in the experiment.

A. Experimental Setting of Drug Monograph

We prepared a data set by selected from two well-known online drug monographs, i.e., RxList (http://www.rxlist.com) and DailyMed (http://dailymed.nlm.nih.gov). These drug monographs are combined to one mixed data set. For two-step classification, data preparations for each step are described as follow.

1) The first step

When we study the therapeutic classes of drug monographs, categorized in AHFS Drug Information 2012. We find out that a drug monograph is often only one therapeutic class. However, there are significant numbers of drug monographs that have more than one class and the maximum number of classes is three. In order to investigate our proposed method in a systematic way, three classes in data collection, i.e., one, two and three categories of drug monographs are used in this step. The total number of drug monographs from each website is 230 monographs. The distribution of the numbers of therapeutic classes of these drug monographs is presented in Table I.

TABLE I: Distributions of the Numbers of Therapeutic Classes for Each Website

The Number of Therapeutic Classes	The Number of Drug Monographs
One	201
Two	24
Three	5

The total number of drug monographs we used is 460 monographs from both websites. The data set for the first step classification is arranged to three classes, i.e., single category, two categories and three categories. The numbers of drug monographs for each category are 402, 48 and 10, respectively.

2) The second step

In this step, each drug monograph is assigned the therapeutic class (classes). There are 11 therapeutic classes for the data collection. The distribution of the numbers of drug monographs for each website is shown in Table II. Note that the total number of drug monographs is not 230 but 264 due to the fact that some monographs contain multiple therapeutic classes. This comes from $(201 \times 1) + (24 \times 2) + (5 \times 3)$ from Table I.

TABLE II: Distributions of the Numbers of Drug Monographs for Each Website

Therapeutic Class	The Number of Drug Monographs
Anti-infective Agents	27
Antineoplastic Agents	26
Autonomic Drugs	24
Blood Formation, Coagulation, and Thrombosis	19
Central Nervous System Agents	31
Cardiovascular Drugs	25
Eye, Ear, Nose and Throat (EENT) Preparations	24
Gastrointestinal Drugs	25
Hormones and Synthetic Substitutes	28
Skin and Mucous Membrane Agents	25
Vitamins	10

All experiments were performed with 5-fold cross

validation. As a preprocessing, some stop words (e.g., a, an, the) are excluded from all data sets. All HTML tags (e.g., <B>, <HTML>) were omitted from documents to eliminate the affect of these common words and typographic words. This may be helpful to make classification processes not depend on any specific format. A unigram model is applied in all experiments. The term weighting, $\overline{tf} \times idf / sd$, is used for prototype vectors on both steps of classification. The default term weighting for a test document is $tf \times idf$. The prototype and test document vectors are normalized by their length. The cosine similarity is used. The value of score is in range from 0 to 1.

The evaluations of the first step, second step and overall are described in the topic of evaluation.

B. Evaluation

The evaluation of text classification uses a confusion matrix that summarize the number of instances predicted correctly or incorrectly by a classification model for each class c_k is shown in Table III. The terminology is described as follow:

- True positive (TP) corresponds to the number of positive examples correctly predicted by the classification model.
- False negative (FN) corresponds to the number of positive examples wrongly predicted as negative by the classification model.
- False positive (FP) corresponds to the number of negative examples wrongly predicted as positive by the classification model.
- True negative (TN) corresponds to the number of negative examples correctly predicted by the classification model.

TABLE III: THE CONFUSION MATRIX FOR A CATEGORY C_k

Category c_i		Predict Class Yes	No
Actual Class	Yes	TP_k	FN_k
	No	FP_k	TN_k

The two widely used metrics employed in applications are successful detection of one of the classes is considered more significant than detection of the other classes. A formal definition of these metrics is given below.

$$Pr_k = \frac{TP_k}{TP_k + FP_k}$$

$$Re_k = \frac{TP_k}{TP_k + FN_k}$$

Precision of a class c_k (Pr_k) determines the fraction of records that actually turns out to be positive in the class c_k that a classifier has declared as a positive class. The higher the precision is, the lower the number of false positive errors committed by the classifier. Recall of a class c_k (Re_k) measures the fraction of positive examples of a class c_k correctly predicted by the classifier. Classifiers with large recall have very few positive examples misclassified as the negative class. Precision and recall can be summarized into another metric known as the F_1-measure that given below.

$$F_{1k} = \frac{2 \times Pr_k \times Re_k}{Pr_k + Re_k}$$

A high value of F_1-measure ensures that both precision and recall are reasonably high.

In this paper, a complete classification which is defined as a monograph is correctly classified if it is correctly classified in both steps. In the first step, monographs are classified into single category, two categories and three categories. For the second step, all monographs will be classified into therapeutic classes. However, the correct classified monographs from the first step will be considered. A ranked list of possible categories is used for assigning a class or classes, according to the number of categories which the classifier assigns to a document in the first step. Evaluation of performance for one, two, three categories, and overall, is measured by classification accuracy (Acc). The equation of classification accuracy is show as follow.

$$Acc = \frac{\text{The number of documents assigned with correct classes}}{\text{The total number of test documents}}$$

VI. EXPERIMENTAL RESULTS

A. The First Step Classification

The number of drug monographs from the two websites we used is 460 monographs. The data set for the first step classification is arranged to three classes, i.e., single category (One), two categories (Two) and three categories (Three). The numbers of drug monographs for each category are 402, 48 and 10, respectively. The confusion matrix of the first step classification is show in Table IV. Using information from Table IV, The values of *Pr*, *Re* and F_1 for each class, are presented in Table V.

TABLE IV: THE CONFUSION MATRIX FOR THE FIRST STEP CLASSIFICATION

	Prediction of Numbers of Classes One	Two	Three
One	392	6	4
Two	0	45	3
Three	0	2	8

TABLE V: THE *PR*, *RE* AND F_1 FOR THE FIRST STEP CLASSIFICATION

	The Numbers of Categories One	Two	Three
Precision	1.00	0.85	0.53
Recall	0.98	0.94	0.80
F_1	0.99	0.89	0.64

For the first step, the classifier with the term weighting of *tfidf/sd*, performs well especially for a single category monographs. Performance on monographs with three categories is medium due to the low precision. From Table IV, small portions of single category and two category monographs are classified as three categories. However, the number of monographs for three categories is tiny compare to the others. This makes the values of precision and F_1 are not competitive to the others. A complete classification is applied in this experiment. Therefore, the monographs which are correctly assigned the numbers of real world classes, are considered in the next step. From this reason, a set of suitable environments should be selected to gain the best result.

B. The Second Step Classification

For the second step, all correct classified monographs from the first step will be classified into therapeutic class (classes). A ranked list of possible categories is used for assigning a class or classes, according to the number of categories which the classifier assigns to a document in the first step. The result of classification accuracies for these three groups is show in Table VI.

TABLE VI: THE ACCURACIES FROM THE FIRST AND SECOND STEPS OF CLASSIFICATION

	The Numbers of Categories		
	One	Two	Three
Total Number of Monographs	402	48	10
Correctly Classified from the Step 1	392	45	8
Completely Classified from the Step 2	370	42	8
Accuracy from the Step 2 (%)	94.39%	93.33%	100.00%
Accuracy from Both Steps (%)	92.04%	87.50%	80.00%

From the result, some observations can be made. Classification accuracies of the three groups of monographs from the step 2 are quite competitive. The accuracies from both steps for single category, two categories, three categories, and overall are 92.04%, 87.50%, 80.00% and 91.30%, respectively. From the experimental results, the two-step classification is an efficient method to classify a multi-label document. The summarization of two-step classification in our experiment is presented in Fig. 2.

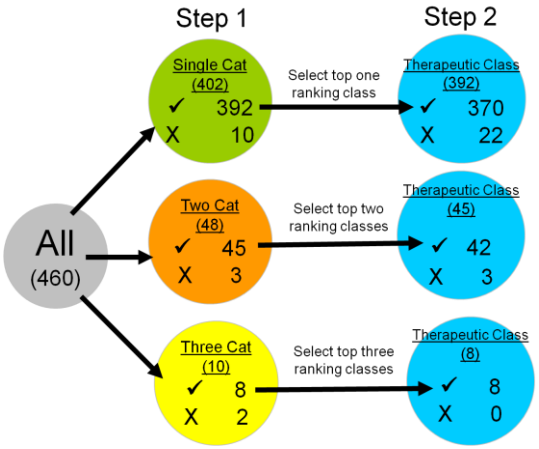

Fig. 2. The summarization of two-step classification for drug monographs.

VII. CONCLUSION AND FUTURE WORKS

In this paper, an algorithm adaptive method called two-step classification was proposed for multi-class and multi-label classification. This method should be applied on a data set which each document had finite number of categories. The minimum and maximum numbers of categories were defined. The first step was assigned the numbers of categories to each document. The real world categories were assigned to a document in the second step. This would be one or more categories based on the number of categories assigned to the document in the first step. The experiment was done using a mixed data set of drug monographs collected from RxList and DailyMed. A centroid-based classifier was applied in both steps. The complete classification, i.e., a correct assigned on both number of classes and therapeutic classes to a document, was used for evaluation. From the results, the two-step classification was an efficient method to classify a document into one class or multiple classes, based on their content.

In this paper, approach of complete classification was used for evaluation. We plan to analyze the result by partial classification, i.e., a document which are partially correct prediction of therapeutic classes (e.g., 2 of 3 classes are correct) should be considered.

REFERENCES

[1] G. K. McEvoy and E. K. Snow, "AHFS Drug Information 2012," American society of health-system pharmacists, Maryland, 2012.
[2] R. J. Roiger and M. W. Geatz, *Data Mining a Tutorial based primer*, Boston, 2003.
[3] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Comput Surv*, vol. 34, no. 1, pp. 1-47, March 2002.
[4] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Mach Learn - Special issue on information retrieval*, vol. 39, no. 2-3, pp. 103–134, May-June 2000.
[5] M. E. Ruiz and P. Srinivasan, "Hierarchical text categorization using neural networks," *Inform Retrieval*, vol. 5, pp. 87–118, 2002.
[6] M. Kubat and M. Cooperson, Jr., "Voting nearest-neighbor subclassifiers," in *Proc. 17th International Conf. on Machine Learning*, California, 2000, pp. 503–510.
[7] E.-H. Han and G. Karypis, "Centroid-based document classification: Analysis and experimental results," in *Proc. 4th European Conference on Principles of Data Mining and Knowledge*, London, 2000, pp. 424–431.
[8] V. Lertnattee and T. Theeramunkong, "Effect of term distributions on centroid-based text categorization," *Inform Sciences*, vol. 158, pp. 89–115, January 2004.
[9] T. Joachims, *Learning to Classify Text using Support Vector Machines*, Dordrecht, NL: Kluwer Academic Publishers, 2002.
[10] G. Tsoumakas and I. Katakis, "Multi-Label Classification: An Overview," *Int J Data Warehous*, vol. 3, no. 3, pp. 1-13, July–September 2007.
[11] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier. "Chains for Multi-label Classification," *Mach Learn*, vol. 85, no 3, pp. 333-359, 2011.
[12] A. Fujino, H. Isozaki, and J. Suzuki, "Multi-label Text Categorization with Model Combination based on F1-score Maximization," in *Proc. 3rd International Joint Conference on Natural Language Processing*, 2008, pp. 823-828.
[13] L. Hua, "Research on Multi-classification and Multi-label in Text Categorization," in *Proc. International Conference on Intelligent Human-Machine Systems and Cybernetics*. Hangzhou, 2009, vol. 2, pp. 86-89.
[14] W. Cheng and E. Hüllermeier, "Combining instance-based learning and logistic regression for multilabel classification," *Mach Learn*, vol. 76, no.2–3, pp. 211–225, September 2009
[15] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inform Process Manag*, vol. 24, no. 5, pp. 513–523, 1988.
[16] A. Singhal, G. Salton, and C. Buckley, "Length normalization in degraded text collections," in *Proc. 5th Annual Symposium on Document Analysis and Information Retrieval*, 1995, pp. 15-17.

Verayuth Lertnattee received a bachelor degree in Pharmacy, a master degree in Science (Computer Science) from Chulalongkorn University in 1989, 1996, respectively. He also received a master degree in Science (Pharmacy) from Mahidol University in 1991. He received a Ph.D. in Technology from Sirindhorn International Institute of Technology, Thammasat University. His research interests include data mining in medical, herbal, pharmaceutical information, information retrieval and collective intelligence.

Chanisara Lueviphan received a bachelor degree in Pharmacy from Silpakorn University in 2007. She is a master student in Faculty of Pharmacy, Silpakorn University. Her research interests include text mining in medical data and hospital information system.