Multi-label Classification using Logistic Regression Models for NTCIR-7 Patent Mining Task

Akinori Fujino and **Hideki Isozaki**
NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan 619-0237
{a.fujino, isozaki}@cslab.kecl.ntt.co.jp

Abstract

We design a multi-label classification system based on a machine learning approach for the NTCIR-7 Patent Mining Task. In our system, we employ a logistic regression model for each International Patent Classification (IPC) code that determines the IPC code assignment of research papers. The logistic regression models are trained by using patent documents provided by task organizers. To mitigate the overfitting of the logistic regression models to the patent documents, we design the feature vectors of the patent documents with feature weighting and component selection methods utilizing a research paper set. Using a test collection for the Japanese subtask of the NTCIR-7 Patent Mining Task, we confirmed the effectiveness of our multi-label classification system.

Keywords: *machine learning, logistic regression model, different training and test distributions, component selection, feature weighting.*

1 Introduction

The goal of the NTCIR-7 Patent Mining Task [7] is to develop multi-label classification systems that assign multiple International Patent Classification (IPC) codes to research papers. IPC codes were originally developed in order to categorize patent documents and make it possible to search the documents efficiently. However, providing research papers with IPC codes means that we can also retrieve papers dealing with the same technological field as a given patent document. This retrieval procedure is required when deciding whether or not to validate a patent. Therefore, classification systems capable of this task are useful for patent analysts in government patent offices or the intellectual property divisions of private companies.

We design a multi-label classification system based on a machine learning approach for dealing with the NTCIR-7 Patent Mining Task. In our formulation, we assume the independence of IPC codes and design a binary classifier for each IPC code that determines whether or not to assign an IPC code to a research paper. We employ *logistic regression models* [5] as the binary classifiers. The logistic regression model designed for each IPC code provides the probability of assigning an IPC code to a research paper. Using the probabilities given by the logistic regression models, our system ranks IPC codes for research papers.

For this task, patent documents assigned with IPC codes and research papers are used as training data for constructing multi-label classification systems, and other research papers are used as test data that require classification. Since the research papers are not assigned with IPC codes, we need to train classification systems by using patent documents. However, research papers are different in form from patent documents, and so the word distribution in the former will be different from that in the latter. Namely, a test distribution is distinct from a training distribution. In such settings, the classification systems may be overfitted in training distributions and thus may not provide good classification performance for test data.

To mitigate this overfitting problem, we design the feature vectors of the training and test examples by using a research paper set. We employ vocabulary words included in the research paper set as features, and weight the feature values of each data examples with inverse document-frequencies of vocabulary words in the research paper set. We also use term frequencies only in the *title* and *abstract* components of patent documents to provide their feature vectors, because other components such as *claim* and *specification* are not included in research papers. Using a test collection provided by the NTCIR-7 organizers, we show the effectiveness of our multi-label classification system and the effect of our feature weighting and component selection methods utilizing a research paper set.

2 Multi-label Classification System based on Logistic Regression Models

For the NTCIR-7 Patent Mining Task, we design a multi-label classification system that assigns research papers with IPC codes. Let the feature vector of a research paper be denoted by $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_V)^T$ and a class be represented by an IPC code assignment vector $\boldsymbol{y} = (y_1, \ldots, y_k, \ldots, y_K)^T$, $y_k \in \{1, -1\}$, where $y_k = 1$ $(y_k = -1)$ in indicates that the research paper $\boldsymbol{x}$ is assigned (unassigned) with the kth IPC code. K is the total number of IPC codes, and $\boldsymbol{a}^T$ represents the transposed vector of $\boldsymbol{a}$.

In our multi-label classification system, we design K logistic regression models each of which provides the probability $P(y_k = 1|\boldsymbol{x})$ of assigning a research paper $\boldsymbol{x}$ with the kth IPC codes, using training dataset $D = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$. We use patent documents assigned with IPC codes as the training dataset for the logistic regression models. Then, we rank the top 1000 IPC codes for a research paper $\boldsymbol{x}$ based on the values of $\{P(y_k = 1|\boldsymbol{x})\}_{k=1}^K$ provided by the logistic regression models. In this section, we review the logistic regression model we used to construct our system, and describe our method for designing features of patent documents and research papers.

2.1 Logistic Regression Models

A binary classifier based on a logistic regression model learns the mapping of a feature vector $\boldsymbol{x}$ to a category label assignment y_k for the kth category label by modeling conditional probability $P(y_k|\boldsymbol{x})$ directly. The conditional probability is modeled as

$$P(y_k|\boldsymbol{x};\boldsymbol{\theta}_k) = \frac{1}{1+\exp(-y_k\boldsymbol{\theta}_k^T\boldsymbol{x})}, \quad (1)$$

where $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{ki}, \ldots, \theta_{kV})^T$ is a model parameter vector.

Using training dataset $D = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$, we can estimate the $\boldsymbol{\theta}_k$ value to maximize the conditional log-likelihood with respect to the training data and the logarithm of a prior probability distribution $p(\boldsymbol{\theta}_k)$ as

$$J(\boldsymbol{\theta}_k) = \sum_{n=1}^{N} \log P(y_{nk}|\boldsymbol{x}_n;\boldsymbol{\theta}_k) + \log p(\boldsymbol{\theta}_k). \quad (2)$$

We use a Gaussian prior [4] as

$$p(\boldsymbol{\theta}_k) \propto \prod_{i=1}^{V} \exp\left(-\frac{\theta_{ki}^2}{\sigma^2}\right), \quad (3)$$

where σ is the hyperparameter of the Gaussian prior and its value should be set for training the logistic regression model. We compute the estimate of $\boldsymbol{\theta}_k$ to maximize $J(\boldsymbol{\theta}_k)$ by using the L-BFGS algorithm [6], which is a quasi-Newton method. In this computation, global convergence is guaranteed, since $J(\boldsymbol{\theta}_k)$ is a concave function of $\boldsymbol{\theta}_k$.

2.2 Feature Design Method

In the NTCIR-7 Patent Mining Task, patent documents assigned with IPC codes are employed as training data, while research papers are used as the target data that should be assigned with IPC codes by the classification systems. The vocabulary words that play an important role in the classification of research papers may be different from those of patent documents. To mitigate the overfitting of our system to patent documents, we design features by utilizing the research paper set provided as training data by the task organizers. We expect to obtain a more suitable classification system for the target data by designing features. In this section, we describe feature weighting and component selection methods used for the feature design.

2.2.1 Feature Weighting

For the feature weighting of the training and test examples, we utilize the document frequencies of the vocabulary words in the research paper set that we used as training data. Let $\boldsymbol{x}_s = (x_{s1}, \ldots, x_{si}, \ldots, x_{sV})^T$ be the feature vector of the sth data example, where V is the total number of vocabulary words. The feature value x_{si} with respect to the ith vocabulary word is computed as

$$x_{si} = \frac{tf(s,i)\cdot w(i)}{Z_s}, \quad (4)$$

where $Z_s = \sum_{i=1}^V tf(s,i)\cdot w(i)$, and $tf(s,i)$ is the term frequency of the ith vocabulary word included in the sth data example. $w(i)$ is the weight designed by using the document frequency of the ith vocabulary word in the research paper set, and is computed as

$$w(i) = \begin{cases} \log \frac{M}{df(i)} & \text{as } df(i) \neq 0 \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

where M represents the total number of research papers, and $df(i)$ is the number of research papers containing the ith vocabulary word.

2.2.2 Component Selection

The patent documents we used as training dataset $D = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$ for our system have multiple components such as *title*, *abstract*, *claim*, and *specification*. However, the claim and specification components are characteristic of patent documents and are not included in research papers. To mitigate the overfitting of our system to the claim and specification

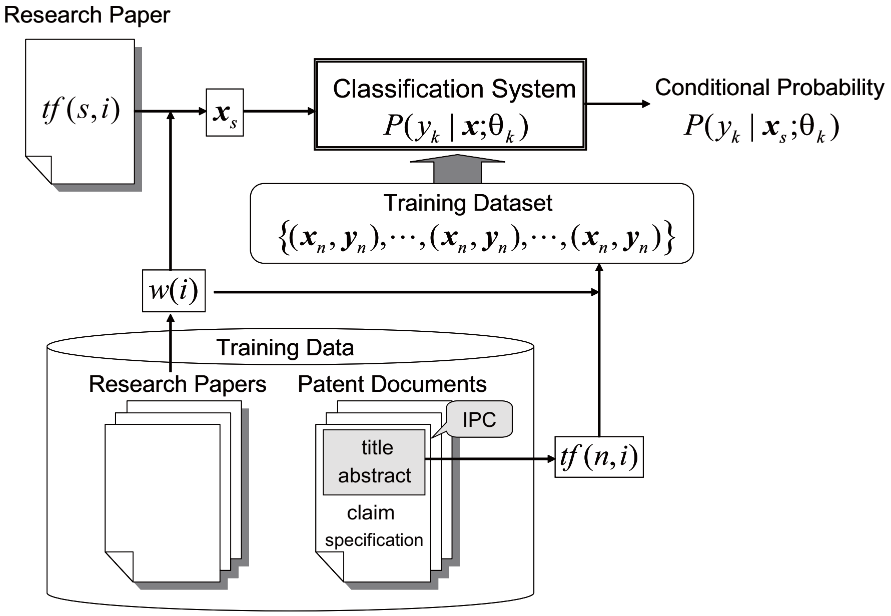
Research Paper
$tf(s,i)$
$\boldsymbol{x}_s$
Classification System
$P(y_k \mid \boldsymbol{x};\theta_k)$
Conditional Probability
$P(y_k \mid \boldsymbol{x}_s;\theta_k)$
Training Dataset
$\{(\boldsymbol{x}_n,\boldsymbol{y}_n),\cdots,(\boldsymbol{x}_n,\boldsymbol{y}_n),\cdots,(\boldsymbol{x}_n,\boldsymbol{y}_n)\}$
$w(i)$
Training Data
Research Papers
Patent Documents
IPC
title
abstract
claim
specification
$tf(n,i)$

Figure 1. Outline of classification system

components of patent documents, we use only title and abstract as components for designing the feature vectors of patent documents. Namely, we provide the $tf(n,i)$ value of the nth patent document by counting the number of the ith vocabulary word included in the title and abstract components.

Figure 1 shows an outline of our system, which employs the feature weighting and component selection methods.

3 Experiments

3.1 Test Collections

We examined the performance of our multi-label classification system for the Japanese subtask of the NTCIR-7 Patent Mining Task. For this subtask, Japanese patent documents published by the Japanese patent office from 1993 to 2002 were given to us for training multi-label classification systems that assign Japanese research papers with IPC codes. The abstracts of Japanese research papers included in the NTCIR-1 and -2 CLIR task text collections were also given to us as training data. The abstracts of 879 Japanese research papers were selected as test examples for the formal run of this subtask by the NTCIR-7 organizers. We utilized the Japanese research papers included in the NTCIR-1 and -2 test collections for feature weighting detailed in Section 2.2, and trained logistic regression models in our system using the patent documents.

We extracted nouns, verbs, and adjectives from patent documents and research papers by using MeCab[1] and utilized these words to provide the feature vectors of these data examples. Vocabulary words included in only one patent document or research paper were removed from the feature vectors.

3.2 Evaluation Results

With the standard TREC-style evaluation method, we calculated recall and precision as regards an IPC code ranking for each research paper, and we summarized these scores in terms of the mean average precision (MAP). These evaluation scores were averaged over all the research papers provided as test data.

Table 1 shows the recall-precision curve and MAP obtained with our multi-label classification system. The evaluation scores in the table were examined by using “trec_eval.prl,” which was provided by the task organizers. The *CS+FW* column in the table shows the performance of our system when providing the feature vectors of training and test examples with the feature design method described in Section 2.2. The feature design method consists of component selection (CS) for patent documents and feature weighting (FW). We also examined the performance of our system when employing either component selection or feature weighting. The experimental results are shown in the *CS* and *FW* columns in the table. The MAP-scores for the CS+FW and FW settings in the table were evaluated as the official results for our system by the task organizers.

[1] http://mecab.sourceforge.net/

Table 1. Recall-precision curve and MAP with multi-label classification system based on logistic regression models

Recall	Precision		
	CS+FW	CS	FW
0.0	**0.5331**	0.5254	0.4641
0.1	**0.5331**	0.5254	0.4641
0.2	**0.5307**	0.5222	0.4613
0.3	**0.4996**	0.4869	0.4233
0.4	**0.4562**	0.4464	0.3734
0.5	**0.4452**	0.4363	0.3667
0.6	**0.3456**	0.3294	0.2728
0.7	**0.2980**	0.2868	0.2380
0.8	**0.2828**	0.2730	0.2262
0.9	**0.2758**	0.2663	0.2223
1.0	**0.2757**	0.2661	0.2223
MAP	**0.3964**	0.3855	0.3303
Run ID	NTTCS4	-	NTTCS1

In the FW settings, all the components of the patent documents were used to provide feature vectors. However, a large number of vocabulary words were included in all the components of each patent document. When using all the vocabulary words to design the feature vectors of the patent documents, we require a very long processing time to train logistic regression models. To reduce the computational costs incurred by training, we designed the feature vector of the sth patent document by using vocabulary words whose x_{si} values, given by Eq. (4), were large, where we selected the largest number of the vocabulary words whose sum of x_{si} was less than 0.995.

As shown in Table 1, our system provided better classification performance in the CS+FW setting than in the CS and FW settings. We confirmed that both the component selection of the patent documents and the feature weighting were useful for improving the classification performance. Our system performed better in the CS setting than in the FW setting. This result indicates that designing the feature vectors of the patent documents by using only their title and abstract was effective as regards obtaining a good classification performance in the Japanese subtask of the NTCIR-7 Patent Mining Task.

4 Conclusion

We designed a multi-label classification system based on logistic regression models for the NTCIR-7 Patent Mining Task. A logistic regression model for each IPC code was used for determining the IPC code assignment of research papers. We used patent documents for training the logistic regression models and designed the feature vectors with feature weighting and component selection methods utilizing a research paper set. Using a test collection provided by the NTCIR-7 organizers, we confirmed experimentally the effect of our feature design method on the classification performance of our multi-label classification system for the Japanese subtask of the NTCIR-7 Patent Mining Task.

Future work will involve applying domain adaptation [1, 2] and semi-supervised learning [8, 3] methods to developing classification systems trained by using both patent documents and research papers as labeled and unlabeled examples.

Acknowledgement

This work is partly supported by MEXT Grant-In-Aid for Scientific Research of Priority Areas: Cyber Infrastructure for the Information-Explosion Era.

References

[1] S. Bickel, M. Bruckner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning (ICML-2007)*, pages 81–88, 2007.

[2] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems 20*, pages 129–136. MIT Press, Cambridge, MA, 2008.

[3] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-supervised Learning*. MIT Press, Cambridge, MA, 2006.

[4] S. F. Chen and R. Rosenfeld. A Gaussian prior for smoothing maximum entropy models. Technical report, Carnegie Mellon University, 1999.

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York Berlin Heidelberg, 2001.

[6] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Programming, Ser. B*, 45(3):503–528, 1989.

[7] H. Nanba, A. Fujii, M. Iwayama, and T. Hashimoto. Overview of the patent mining task at the NTCIR-7 workshop. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, 2008.

[8] X. Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin, 2005.