Semi-Automatic Labeling of Training Data Sets in Text Classification

Nayereh Ghahreman
Department of Computer Engineering, University of Isfahan, Iran
E-mail: n.ghahraman@gmail.com

Ahmad Baraani Dastjerdi
Department of Computer Engineering, University of Isfahan, Iran
Tel: 98-311-793-4095 E-mail: ahmadb@eng.ui.ac.ir

Received: May 3, 2011 Accepted: September 9, 2011 Published: November 1, 2011
doi:10.5539/cis.v4n6p48 URL: http://dx.doi.org/10.5539/cis.v4n6p48

Abstract

Web includes digital libraries and billions of text documents. A fast and simple search through this sizeable set is important for users and researchers. Since manual or rule based document classification is a difficult, time consuming process, automatic classification systems are absolutely needed. Automatic text classification systems demand extensive and proper training data sets. To provide these data sets, usually, numerous unlabeled documents are labeled manually by experts. Manual labeling of documents is a difficult and time consuming process. Moreover, in manual labeling, due to human exhaustion and carelessness, there is the possibility of mistakes.

In this study, semi-automatic creation of training data set has been proposed in a way that only a small percentage of this extensive set's documents is labeled manually and the remaining percentage is done automatically. Results show that by labeling only ten percent of the training set, remaining documents can be automatically labeled with 98 percent of accuracy. It is worth mentioning that this reduction in accuracy only occurs in standard data sets, while for large practical data sets, this reduction is trivial compared to the accuracy reduction resulted by human exhaustion and carelessness.

Keywords: Semi-supervised learning, Supervised learning, Test set, Text classification, Training set

1. Introduction

Automatic text classification is a wide area of research in the literature. One technique proposed for text classification is the machine learning solution. A promising solution in machine learning towards achieving this goal is Support Vector Machine (SVM), although other solutions have been implemented on Reuters standard data set.

Reference (Sebastiani, 2002) offers a brief survey on text classification. That paper introduces various techniques for text classification with a focus on machine learning solutions. One of these techniques is SVM which its results expose it as a promising technique for text classification (Dumais, Platt, Heckerman, & Sahami, 1998; Joachims, 1998).

Dumais et al. evaluated SVM on Reuters 21578 standard data set (Dumais et al., 1998). Their results showed that SVM has the best precision in text classification. This approach also has a high training speed.

Since there are too many text documents and they are expanding every day, a training set including a high number of documents is absolutely essential to classify a new accurate set of documents. In this case, with a precise and rich training of the classifier, classification of large new sets of documents will be done automatically. Thus the first and most important requirement for classifying documents is to prepare a large data set by decreasing the human factor in the process of creating such data set. A task that otherwise would be very difficult for a human expert will be carried out automatically. Moreover, the training data set will be protected from false labeling caused by human exhaustion and recklessness; hence prediction of the classifier system is improved. In this paper, using Support Vector Machine (SVM) and semi-supervised learning, a huge semi-automatic training set is provided for training of the classifier system. It is worth mentioning that the proposed approach eliminates problems like the excessive manual labeling and significant decrease in accuracy.

2. Definitions

2.1 Text Classification

Text classification is an ongoing topic in information retrieval and machine learning literature (Ko, Park & Seo, 2002). The goal of text classification is to classify documents into a certain number of predefined categories so when a new text is arrived it will be assigned to one or more of these categories otherwise it won't belong to any of them (Ko & Seo, 2000).

Text classification is the process of assigning a Boolean value to each pair of $<d_j,c_i> \in D \times C$, in which, D is the set of documents, and C is the set of predefined categories of texts. In this process, a value of T or F is assigned to $<d_j,c_i>$. T presents membership of document d_j to category c_i and F represents no membership. Thus the objective function is defined as below (Sebastiani, 2002):

$$\Phi: D \times C \rightarrow \{T, F\}$$

2.2 Supervised Learning

Supervised learning requires a large amount of labeled data, but the data labeling process can be expensive and time consuming, as it requires the efforts of human experts (Abdel, Hady, Schwenker & Palm, 2009). Furthermore, this process requires human effort and labor. In this method, the training set is labeled manually, and then the classifier system is trained by using the provided training set and prepared to predict the input data in the future.

2.3 Semi-Supervised Learning and Co-Training

In semi-supervised learning, the training set is provided semi-automatically. Co-Training is a semi-supervised learning method that can reduce the amount of required labeled data through exploiting the available unlabeled data to improve the classification accuracy (Abdel et al., 2009).

2.4 Support Vector Machine

SVM is a machine learning model which was introduced by V. N. Vapnik (1995). The idea behind this algorithm is to find an optimum hyperplane to discriminate two classes. The optimum hyperplane is the one which separates two classes, and In addition has the highest margin and distance from the sample of these classes. After finding the optimum hyperplane, the position of test data related to the hyperplane is measured and based on that relative position, its class is determined.

SVM Advantages:

a) It can be used for cases in which the discriminator plane of two classes is not linear.

b) It is a powerful statistical model for cases with very large feature sets.

Recently, this model was used for text classification successfully. T. Joachims performed text classification using SVM and obtained more promising results than other machine learning methods like KNN and Bayes (Burges, 1998). J. T. Kwork (1998) also used SVM to classify the Reuters data set and obtained better results compared to KNN.

3. Text Classification Phases and Their Disadvantages:

Text classification includes two phases described below:

- **Training Phase (or Preprocessing Phase)**
- **Classification Phase**

In the training phase, documents of the training set are labeled using a supervised method. Also, In addition feature extraction and feature selection are carried out in this phase. In classification phase, classification of input documents is done and using automatic classifier systems, new documents are classified (Ko, Park & Seo, 2002).

Labeling the training set and features selection in training phase cause problems which are described below. Labeling the training set which includes a large number of documents is performed manually, which is a difficult and time consuming process (Abdel et al., 2009). In a number of studies some research, to solve this problem unsupervised or semi-supervised learning has been used (Ko, Park & Seo, 2002; Ko & Seo, 2000).since supervised learning requires a large number of labeled data and the task of data labeling can be costing and time consuming, which needs experts' effort and labor (Abdel et al., 2009).

As mentioned before, another fundamental problem in training phase is high dimension feature space in texts. Consequently, feature selection is one of the most important issues in text preprocessing for text classification.

A proper measure for feature selection can lead to an increase in accuracy and speed of text classification. In previous studies, researchers have proposed useful and effective solutions for the problem of high dimensionality of feature space, including proper filters like Term Frequency, Chi-Square (Shang, Huang, Zhu, Lin, Qu & Wang, 2007), Multi-classOdds Ratio (MOR) (Chen, Huang, Tian & Qu, 2009), Class Discriminating Measure (CDM) (Chen et al., 2009), Poisson distribution (Ogura, Amano & Kondo, 2009), Gini index (Ogura, Amano & Kondo, 2009), Information Gain (IG) (Yoon, Lee & Lee, 2006), Multinomial mixture model with feature selection (M3FS) (Li & Zhang, 2008) and other proper filters for decreasing the dimensionality of feature space.

After handling these problems, increasing in accuracy and reducing the time needed for of text classification in applications such as information retrieval, document organization (Shang, Huang, Zhu, Lin, Qu & Wang, 2007) and web page categorization will be achieved.

4. Research Goal

The goal of this research is to deal with decrease the problem of manual labeling of training set in preprocessing step and training phase. Because while providing a set of unlabeled documents is a fairly simple task, manual labeling of those documents in order to create a rich training set is a difficult process (Ko & Seo, 2000).

In this study, using machine learning methods, a few percentages of documents of the training set are labeled manually and the rest is dealt with automatically and unsupervised. Using the resulted training set, which contains a large number of labeled documents and selecting features with previous approaches, accuracy of text classification for the training set is improved or almost the same as before.

In this way, so that both problems of excessive manual labeling of documents, which is a costing and time consuming process and false labeling caused by human exhaustion and recklessness are resolved, thus speed and classification accuracy is increased.

5. Related Works

In previous attempts to maintain and handle the problem of feature selection in text classification, which is an important problem due to high dimensionality of feature space, a number of researches were conducted (Chen et al., 2009; Ogura, Amano & Kondo, 2009; Li & Zhang, 2008; Yoon, Lee & Lee, 2006; Kim, Han, Rim & Myaeng, 2006) and promising results have been reported. However, there is a need for addressing the problem of preparing the training set. Because with a consecutive increase in the number of new documents to be classified, a large training set to increase the classification accuracy of new documents is absolutely essential. Various methods for semi-automatic preparation of a large training set for text classification have been proposed. Below are brief introductions about these methods and arguments on their disadvantages.

In the method proposed in (Ko & Seo, 2000) instead of using the training document set, the training term set has been used. In this solution, a list of each category's relevant words is defined manually. These words define each category's specific features. After the specific words for each category are determined, their synonyms should also be taken into consideration. In this method, available documents are shrunk into terms. Terms which include predefined words for each category are assigned as that category's indicator. Then, categories of non-classified terms are determined using a similarity measure between the terms and indicator terms. These comparisons are carried out by weighting the words and terms. Below, some of these measures have been described.

- To calculate the weight of a term, term frequency (TF) and inverse category frequency (ICF) must be calculated. These measures are defined as below.

TFij: Number of occurrence of the term ti in the jth category

ICF: This measure is defined as below:

$$ICF\ i = Iog(M) - Iog(CF\ i)$$

CFi: Number of categories which contain ti

M: Total number of categories

Based on these measures, wij, i. e. the weight of ti in the jth category is calculated as below:

$$wi\ j = TFij * ICF\ i = TFij * (log(M) - log(CF\ i))$$

Finally, to calculate the weight of a term in the jth category (Wij), the weights of the words inside that term are used:

$$Wi\ j = (w1j+w2j+\ldots+wNj)/N$$

N: Number of the words comprising the term

Then the measure described below which estimates the similarity between non-classified terms and the indicator term of each category is used to classify the non-classified terms.

$$\underset{ci \in C}{\text{sim}(X,ci)} = 1/n \sum_{j=0}^{n} \underset{Sj \in Rci}{\text{sim}(X,Sj)}$$

$$\underset{ci \in C}{\text{sim}(X,ci)} = \max_{Sj \in Rci}\{\text{sim}(X,Sj)\}$$

X: Non-classified term
$C = \{c_1, c_2, \ldots, c_m\}$: Set of categories and classes
$Rci = \{s_1, s,\ldots, s_n\}$: Set of indicator terms of category c_i
Each term is assigned to a category whose indicator terms are more similar to it:

$$\max_{ci \in C}\{\text{sim}(X, ci)\} \geq \mu + \theta\sigma$$

Using statistical measures a document is assigned to a category which contains the most terms belonging to that category.

Other approaches based on dividing documents into terms and using various measures for term classification have been proposed including using text summarization methods in term classification. These methods try to label documents of the training set based on important terms extracted from documents using term importance evaluation measures. This approach also requires manual determination of important terms (Ko, Park & Seo, 2002).

In another solution (Chen et al., 2009) training set is added using semi-supervised learning. In above solution a semi-supervised learning algorithm based on graph theory is introduced and used to determine positive learning samples. These samples are used to add a training set. In the evaluation phase, the effect of increasing new training samples on text classification is determined by using SVM. This semi-supervised learning algorithm is based on theory of Gaussian random fields in which labels of labeled training samples are propagated over unlabeled data based on probability. To perform this label propagation, a weighted graph can be built. The graph's vertices represent text documents and the weight of each edge represents the similarity between two documents related to the corresponding edge. When two documents are defined as vectors V_i, V_j in the vector space, the weight of each edge is defined as below:

$$w_{ij} = \exp\left(\frac{1}{0.03}\left(1 - \frac{v_i' v_j}{|v_i| \times |v_j|}\right)\right)$$

Wij is the weight of an edge which connects vertex i to vertex j. This weight is used to represent the similarity between two text documents in vector space. Besides this graph, a learning algorithm is also needed to label positive and negative cases. The data set used in this study is a subset of MEDLINE data set. It contains four classes of A, E, G and T.500 training documents which are not belonging to any of those classes are chosen and labeled as negative cases. These negative cases are combined with a set of positive cases that are manually labeled for each category (338, 81, 462 and 36 positive cases for categories A, E, G and T respectively). Then the label propagation process is performed on remaining documents.

One problem with the above method is a significant decrease in the accuracy of text classification for the test set. This proves that semi-automatically provided training set lacks enough accuracy and precision. Another important problem is that a large number of negative documents are labeled manually (35% of documents are labeled manually). Negative document labeling takes a large amount of time and labor since a document is labeled as negative when it does not belong to any of the classes.

6. Proposed Solution

After studying previous approaches it was observed that SVM is the most common method for text classification. Based on this observation, a solution based on the SVM evaluation system is described. The system's task is to prepare to train sets for text classification.

This approach uses co-training algorithms combined with the similarity evaluation measure in order to be able to perform more accurate labeling. The basic selected algorithm is as below:

Training Set Preparation Algorithm:

1. Input:
Set L, a collection of manually labeled documents

Set U, a collection of unlabeled documents from the training set

S, SVM classifier system.

2. Training classifier system S using L
3. Classifying U using the S classifier system:
Selecting positive samples:

The number of these samples in each step is at most 15 percent of the number of L's members which are selected as below.

Samples that their predicted labels are more than 0.9 are selected. If this number was less than one sixth of the desired number a value of 0.1 is subtracted of the threshold value (0.9-0.1 = 0.8). This process continues until at least one sixth of 15 percent or at most 15 percent of L's members are selected as positive samples.

Selecting negative samples:

The number of these samples in each step is at most 15 percent of the number of L's members which are selected as below.

Samples that their predicted labels are less than -0.9 are selected. If this number was less than one sixth of the desired number a value of 0.1 is added to the threshold value (-0.9 + 0.1 = -0.8). This process continues until at least one sixth of 15 percent or at most 15 percent of L's members are selected as negative samples.

Ten features from each category are selected as best features using Term Frequency as a feature selection measure. These are listed in set W. Then the similarity between the selected documents and manually labeled documents is calculated as below.

$$Sim(d_1, d_s) = \sum_{w_i \in W} p(w_i, d_1).p(w_i, d_s).[1 - \log(df(w_i) + 1) / \log(n+1)]^2$$

d1: A manually labeled document

ds: Selected document

p(wi,d): Frequency of the word wi divided to the total number of the words within the document d

df(wi): Total number of manually labeled documents that include the word wi

n: Total number of manually labeled documents.

1. Group of selected documents which the type of their labels is the same as the similar document's are added to L and omitted from U.
2. If U has two members go to step 2.
3. Output: L, the training set which is provided semi-automatically.

Using this algorithm, the research goal was achieved.

In the first step of this algorithm, using a group of labeled documents L, the classifier system is trained. Then the collection of unlabeled documents U is given to the classifier system for labeling. In each step at least one third of 30 percent, and at most 30 percent of L's total number of members are selected as positive and negative samples and added to the collection of labeled documents. This process continues until all the unlabeled documents of the training set are labeled.

It should be noted that the selected threshold value and number of selected documents of the algorithm have been decided based on comparing different states and choosing the most optimal state. In each step, documents are selected, which are farther from the support vector and thus the class predictions for them include fewer errors. These documents are added to labeled documents if they pass the similarity evaluation filter.

Moreover, in this algorithm optimization of feature values is addressed. To fulfill this goal, feature values were normalized, i.e. instead of considering the number of repeated words in the text, the ratio of the number of repeated words in the document to the total number of words was used. By using the normalization a better support vector is produced and in each step by adding new documents, vectors get updated and optimized.

7. General Scheme of the Solution

Phase I: Manual labeling of a few percentage of the training set. This is illustrated more in Figure 1.

Phase II: automatic labeling of the remaining percentage of the training set using classifier system from phase I (Figure 2).

Phase III: Selected documents are added to the collection of labeled documents and deleted from the collection of unlabeled documents. Then if the unlabeled collection still has a member, previous phases repeat.

Experimental Results:

The proposed algorithm is implemented using the MATLAB 7.6, library of libsvm and rcv data set. Experimental results are presented in Table 1. It should be noted that the decrease in accuracy has occurred for the standard data set, while for large practical data sets, this reduction is trivial compared to the accuracy reduction resulted by human exhaustion and carelessness.

Finally, the proposed method was compared with the Automatic Text Categorization using the Importance of Sentences (Ko, Park & Seo, 2002) in Table 2.

8. Conclusion

Web includes digital libraries and billions of text documents. A fast and simple search through this sizeable set is important for users and researchers. Since manual or rule based document classification is a difficult, time consuming process, automatic classification systems are absolutely needed. Automatic text classification systems demand extensive and proper training data sets. To provide these data sets, usually, numerous unlabeled documents are labeled manually by experts. Manual labeling of documents is a difficult and time consuming process. Moreover, in manual labeling, due to human exhaustion and carelessness, there is the possibility of mistakes.

The goal of this research is to deal with decrease the problem of manual labeling of training set in preprocessing step and training phase. In this study, using machine learning methods, a few percentages of documents of the training set are labeled manually and the rest is dealt with automatically and unsupervised. Using the resulted training set, which contains a large number of labeled documents and selecting features with previous approaches, accuracy of text classification for the training set is improved or almost the same as before.

Results show that by labeling only ten percent of the training set, remaining documents can be automatically labeled with 98 percent of accuracy. It is worth mentioning that this reduction in accuracy only occurs in standard data sets, while for large practical data sets, this reduction is trivial compared to the accuracy reduction resulted by human exhaustion and carelessness.

References

Abdel, M., Hady, F., Schwenker & Palm, G. (2009). Semi-supervised learning for tree-structured ensembles of RBF networks with Co-Training, *Neural Networks*.

Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2), 955-974. http://dx.doi.org/10.1023/A:1009715923555

Chen, J., Huang, H., Tian, S. & Qu, Y. (2009) .Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, 36, 432-435.

Dumais, S. T., Platt, J., Heckerman, D. & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management, 148-155.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. *In Proceedings of ECML-98, 10th European Conference on Machine Learning*, 137-142.

Kim, S., Han K., Rim, H. & Myaeng, S. (2006). Some effective techniques for Naïve Bayes text classification. *IEEE Transactions on Knowledge and Data Engineering*, 18, 1457-1466.

Ko, Y., & Seo, J. (2000). Automatic Text Categorization by unsupervised Learning. *In proceedings of the 18th International Conference on Computational Linguistics*, 453-459.

Ko, Y., Park, J. & Seo, J. (2002). Automatic Text Categorization using the Importance of Sentences. In proceedings of the 19th international conference on Computational linguistics International, 1, 1-7.

Kwok, J.T. (1998). Automated text categorization using support vector machine. In Proceedings of the International Conference on Neural Information Processing, Kitakyushu, Japan, Oct., 347-351.

Li, M. & Zhang, L. (2008). Multinomial mixture model with feature selection for text clustering. *Knowledge-Based Systems*, 21, 704-708. http://dx.doi.org/10.1016/j.knosys.2008.03.025

Ogura, H., Amano, H. & Kondo, M. (2009). Feature selection with a measure of deviations from Poisson in text categorization. *Expert Systems with Applications*, 36, 6826-6832. http://dx.doi.org/10.1016/j.eswa.2008.08.006

Reuters-21578 collection. [Online] Available: http://www.research.att.com/~lewis/reuters21578.htm

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys,* 34(1), 1-47. http://dx.doi.org/10.1145/505282.505283

Shang, W., Huang, H., Zhu, Lin, Y., Qu, Y. & Wang, Z. (2007). A novel feature selection algorithm for text categorization. *Expert Systems with Applications*, 33, 1-5. http://dx.doi.org/10.1016/j.eswa.2006.04.001

Vapnik, V. N. (1995). The nature of Statistical Learning Theory. *Springer*.

Yoon, Y., Lee, C., & Lee, G. (2006). An effective procedure for constructing a hierarchical text classification system. *Journal of the American Society for Information Science and Technology*, 57, 431-442. http://dx.doi.org/10.1002/asi.20281

Table 1. Results obtained from rcv_train data set, using libsvm for implementation

Algorithm	**Percentage of manually labeled documents**	**Accuracy for semi-automatic training set**
Section 6 Algorithm	5%	96%±1
Section 6 Algorithm	10%	98%±1
Section 6 Algorithm without the similarity measure (step 3 of the algorithm)	5%	90%±1
Section 6 Algorithm without the similarity measure (step 3 of the algorithm)	10%	96%±1
Section 6 Algorithm, using Term Frequency (TF) as feature value	5%	86%±1
Section 6 Algorithm, using TF divided to the total number of the words inside the document (P) as feature value	5%	96%±1

Table 2. The comparison of the proposed method and the Automatic Text Categorization using the Importance of Sentences

data set	machine learning model	method	F-score
English newsgroup data set	SVM	proposed method	94
English newsgroup data set	SVM	Automatic Text Categorization using the Importance of Sentences (Ko ,Y., Park , J., & Seo ,J., 2002)	86.35
English newsgroup data set	*k*-NN	Automatic Text Categorization using the Importance of Sentences (Ko ,Y., Park , J., & Seo ,J., 2002)	82.6
English newsgroup data set	Naïve Bayes	Automatic Text Categorization using the Importance of Sentences (Ko ,Y., Park , J., & Seo ,J., 2002)	84.35

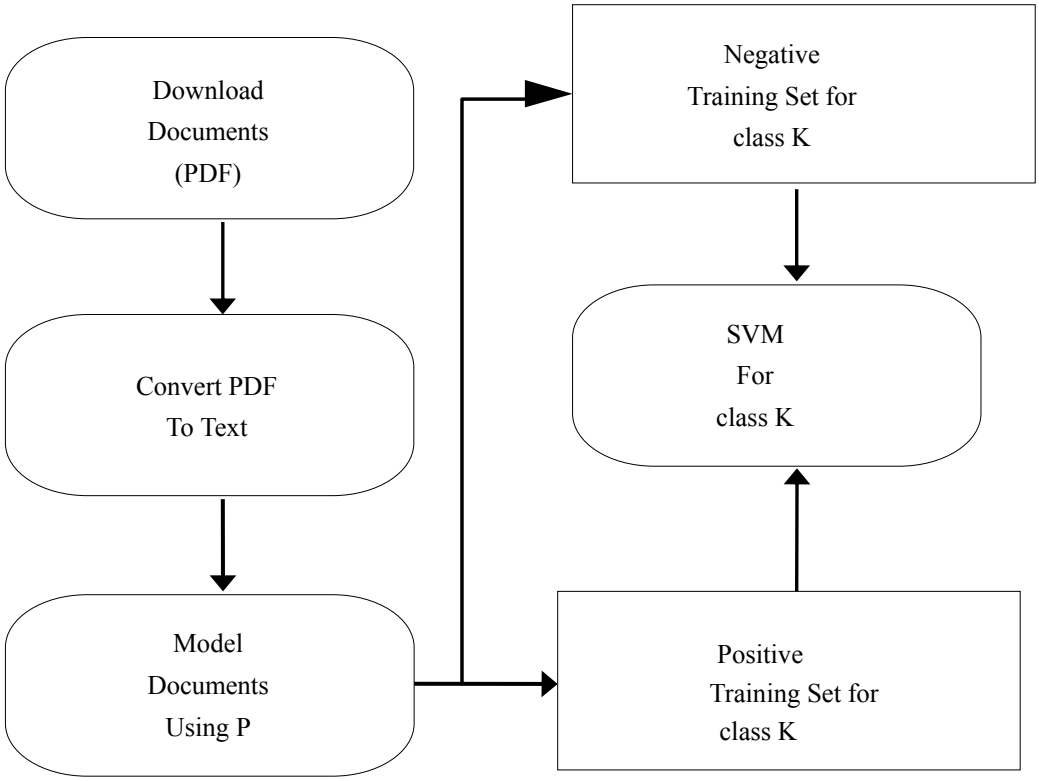

Figure 1. Manual labeling

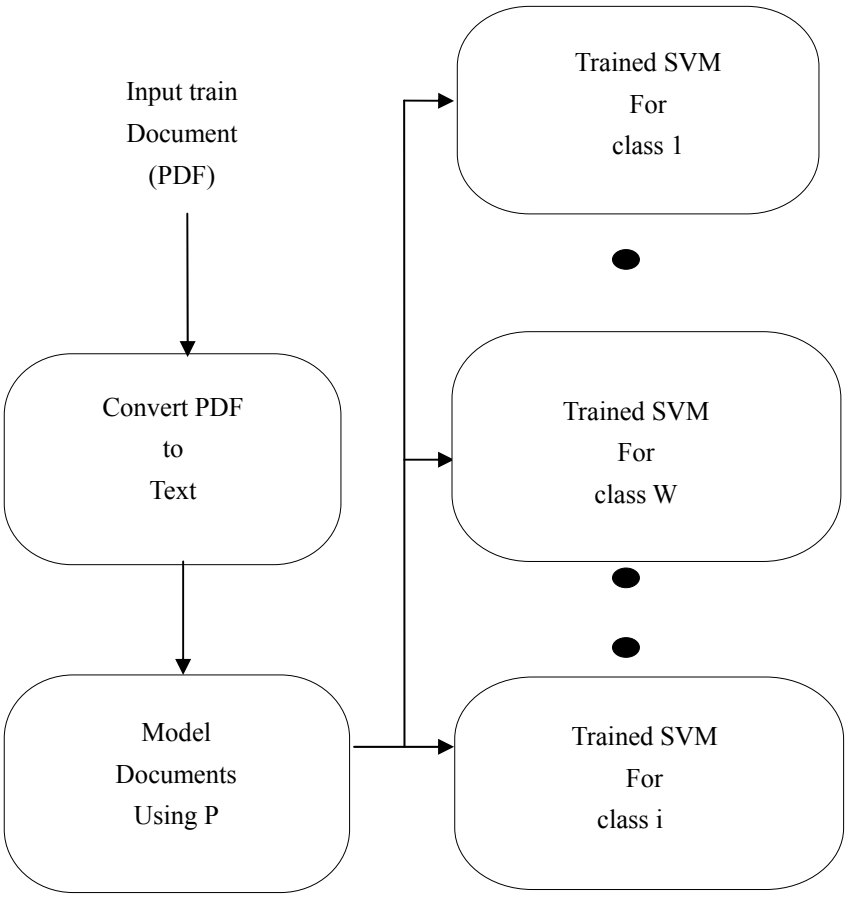

Figure 2. Automatic labeling

ISSN 1913-8989 E-ISSN 1913-8997