Support Vector Machine Based Educational Resources Classification

Tian Xia

***Abstract*—China universities are building informational applications nowadays. Much data are generated anytime and the amount of the data is approaching big data level.**

Universities also build public data platform/data center to integrate and synchronize data from many application systems for data outlining and big data analysis.

However, there are many unstructured raw data that cannot be synchronize with public data platform, as amount of them are huge. Also, all the data, never the less synchronized or unsynchronized, can only shared within single university.

To share data between universities, a rapid way is to classify the data of universities from their similar application systems via Internet. So, it is necessary to extract data through internet and classify these data automatically.

Since the data of Educational Resources has their own particularities that must be taken into consideration in classification, this paper put forward a Support Vector Machine (SVM) based education resources automatic classifier that address this issue.

***Index Terms*—Support vector machine, SVM, education, classification, natural language processing.**

I. Introduction

Recent years, universities in China apply many informational applications or systems to digitalize education, research and administration, such as paperless OA.

In such process, much digital information was generated like information exploration and was collected together by a database system, often called public data platform. Such kind of process keeps data update in its central database. Engineers in information office often try their best to make full use of the data via Big Data technologies. Such data analysis always focuses on role-based data, such as student learning data analysis, teacher data analysis *et al.* However, such role-based data relies in not only the center database, but also the data of application systems of departments, which are not included in the synchronization process of the public data platform. For instance, to analyze teaching process data of a course, it is necessary to study course plan, attendance data, courseware data, course resource data, and interactive data with students, etc. These data belong to many application systems besides the center database, such as the educational administration system, course center *et al.*

Therefore, the data discussed above are not only structured but also unstructured. The method put forward in this paper can apply to the scenario containing the both kinds of data, but mainly focus on the latter one, because the classification results of the latter one is structured data, which can be processed by structured data analysis.

Also, take educational resources for example, the data of educational resources are unstructured raw data. Due to none such data transfer between universities, they become Information Islands which can only be used within a single university, sometimes only one course.

To resolve this issue, researchers tried to establish educational standards to formalize the educational resources for retrieving and sharing educational resources data among universities. However, it is not accomplishable enough to categorize them automatically according to the educational resources standards.

Therefore, currently, in order to share educational resources in an efficient way, it is necessary to develop a technical method to automatically classify the unstructured educational raw data, such as educational resources.

In this paper, a Support Vector Machine (SVM) [1] based education resources automatic classifier is described to address this issue in education resources classifying for example. This method can be applied to other raw educational data, which contains sematic context expressing the main content or purpose of the data.

II. Particularity of Educational Resources

To study educational resource data classification problem, it is better to research on the Particularity of it.

A. Educational Resource Is Specialty-Sensitive

Teachers always focus on the topics of their own teaching and research and publish related educational resources on their websites or course center. So for this reason, the context of Educational Resource often contains many specialty domain terms, which are hard to be correctly Chinese segmented and, as a consequent, difficult to assign a proper weight for such terms.

B. Educational Resource Is Leveled

Educational Resource is always only suitable to certain level of students, who has similar specialty and grade. Therefore, it is better to classify the Educational Resource into categories according to not only specialties but also grades.

C. Educational Standards Vary

There are not very common official educational standards. Therefore, the method must be flexible so that it can adapt the frequent changes of educational standards.

D. Educational Resources Often Contain Hypertext and

Manuscript received May 3, 2015; revised July 21, 2015.

Tian Xia is with Shanghai Second Polytechnic University, Shanghai, 201209, China (e-mail: xiatian@sspu.edu.cn).

Hypermedia

The SVM method cannot process any other kinds of data except plain text. Therefore, it is necessary to pre-process the data to make it suitable to be processed by the method in this paper.

E. The expression of Educational Resources Is Always Normative

Normative expression is suitable for the method in this paper to classify educational resources.

III. Support Vector Machine

Support vector machines (SVM) are based on the principle of Structural Risk Minimization from statistical learning theory. It is a well-known learning algorithm that has been widely used in many applications including classification, estimation and tracking as in [1]-[3] and [4]. SVM finds the closest data vectors called support vectors (SV), to the decision boundary in the training set and it classifies a given new test vector by using only these closest data vectors [5], [6]. The idea of structural risk minimization is to find a hypothesis h for which we can guarantee the lowest true error. In the presence of noise, the idea of using a soft margin was introduced by Vapnik [5].

In this paper, we considers the typical web classification problem with a document collection *D* which is extracted from webpages of course center, in the form of $\{x_i, y_i\}_{i=1}^{l}$, where $x_i \in X \subset R^n$ is the ith sample and $y_i \in \{1,-1\}$ is the corresponding class label. Also, $x^{(j)}$ denotes the jth feature of vector x, hence $x_i^{(j)}$ is the the jth feature of ith sample.

For the given document collection *D*, SVM maps the data point $x_i \in X$ into a high (possibly infinite) dimensional feature space *H* using a nonlinear mapping function $\Phi: X \to H$. The decision boundary of the binary classification problem takes the form of an optimal separating hyper plane $w \cdot \Phi(x) + b = 0$, which is determined by a weight vector w and a bias b. This hyper plane can be obtained by solving the following optimization problem.

$$\begin{aligned} &\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\ &s.t. y_i(w \cdot \Phi(x_i) + b) + \xi_i \geq 1, \\ &\xi_i \geq 0, \quad for \quad i = 1,2,...,l \end{aligned} \tag{1}$$

where ξ_i for $i = 1,2,...,l$ are slack variables introduced to handle the non-separable case. The constant *C*>0 is the penalty parameter that controls the trade-off between the separation margin and the number of training errors, with higher value of *C* focusing more on minimizing error. Using the Lagrange multiplier method, one can easily obtain the following Wolfe dual form of the primal quadratic programming problem:

$$\begin{aligned} &\min_{\alpha_i, i=1,2,...,l} \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j k(x_i, x_j) - \sum_{i=1}^{l}\alpha_i, \\ &subject\ to\ 0 \leq \alpha_i \leq C, i = 1,...,l, \\ &y^T\alpha = 0. \end{aligned} \tag{2}$$

SVM works in the feature space via some nonlinear mapping function $\Phi: X \to H$, which can be defined implicitly by a kernel function $k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$. At the optimal point for (2), either $\alpha_i = 0$, $0 \leq \alpha_i \leq C$ or $\alpha_i = C$. The input vectors for which $\alpha_i > 0$, are termed as support vectors. These are the only important information from the perspective of classification, as they define the decision boundary, while the rest of the inputs may be ignored. For a binary classification problem, the decision function of SVM takes the form [5]:

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{N_S}\alpha_i k(x_i, x) + b), \tag{3}$$

where α_i is corresponding weight of support vector x_i , x is the input pattern to be classified, N_S is the number of support vectors and b is the bias.

IV. Support Vector Machine Based Education Resources Automatic Classifier

The Support Vector Machine based Education Resources Automatic Classifier consists of trainer and classifier.

Trainer analyses the training corpus, which contains extracted text of webpages classified by categories. The classifier analyses the test document, calculate its results of the decision function of SVM and announce the most similar one is the result category.

The common process for trainer and classifier is to formulize the text. Steps are listed below.

A. Pre-process

At the very beginning, text of webpages of course center is extracted by to generate documents in the training corpus and for the test document.

Also, to build document representation, for Chinese, it is necessary to first perform Chinese segmentation, word tagging, etc. Then, documents are segmented to word sequences with part of speech attached.

B. Feature Weighting

Each word in each document will be assigned a weight value to present its sematic importance in the document during this step. Therefore, the word sequences are broken and represented by document vectors, which indicate each words importance in the document.

To get accurate weight value, term weight algorithms play an important role in such process. Many term weight algorithms are put forward. But, TF-IDF [7] is he most widely used one.

TF weight formula is:

$$TF_i = \log_2(tf_{ij}) \tag{4}$$

where tf_i denotes the frequency of term i in document j.

IDF weight formula is:

$$IDF_i = \log_2(\frac{N}{n_j}) + 1 = \log_2(N) - \log_2(n_j) + 1 \tag{5}$$

where N is the total number of documents in the collection and n_j is the number of documents that contain at least one occurrence of the term i.

C. Domain Terms

To emphasize the domain terms, we evaluate their importance concerning specialty and categorize them into 3 levels and assign a domain term coefficient D_i to the TF-IDF algorithm.

$$W = D_i \times TF_i \times IDF_i \tag{6}$$

D. Feature Selection

Features for the text documents are words or phrases appearing in the documents. Feature selection is the process of selecting a subset of much relevant features for use in document vector representation.

The central assumption in feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those providing no more information than selected features, and irrelevant features provide no useful information in any context and may interfere with accurance of classification.

For text representation, each word is considered as a feature. However, this will result in system resources consumption. Also, as it contains redundant information, it also decrease the classification accuracy as well. A careful selection of words is desired instead of all words [8]. A simple unordered list of selected words based on its semantic importance and associated proper weights are usually sufficient to represent a document.

Another reason of feature selection is that a collection of documents is involved rather than individual documents. The main purpose is to make it easy to classify documents. The size of an feature index can be reduced when the stems of words are used instead of all word.

Currently, many methods, such as Mutual information [9], the weight of evidence for text, Information Gain [10], Expected cross entropy [11] *et al.*, have been applied to feature extraction in text classification systems.

Based on our experiment, Information Gain and CHI Square method usually show their better performance and results. In this paper, Information Gain is used and the formula is:

$$\begin{aligned} G(w) = & -\sum_{i=1}^{m} p(c_i)\log p(c_i) \\ & + p(w)\sum_{i=1}^{m} p(c_i \mid w)\log p(c_i \mid w) \\ & + p(\overline{w})\sum_{i=1}^{m} p(c_i \mid \overline{w})\log p(c_i \mid \overline{w}) \end{aligned} \tag{7}$$

where *c* represents categories and w is the weight of terms.

E. Stop Words

Although much redundant information can be removed by the feature selection process, however, noises, generally defined in IR as the insignificant, irrelevant words or stop words, are normally present in any plain text.

Stop words have an average distribution in any standard language corpus and do not normally contribute any information to classification tasks. But, since these stop words have high frequencies of occurrences and the average distribution in documents, they may get higher IDF value than expected.

F. Kernel Function

We build SVM simulation model to classify educational resources. Pretreated data of training documents is to train SVM model. Classification accuracy is achieved through SVM model.

SVM classifier will be trained by the normalized training sample data. The training function is performed by using svmtrain statement which comes from LIBSVM toolbox and sets Kernel function type t. The classification accuracy can be adjusted based on experience through changing the punish parameter c and kernel function parameters g. In this paper, *c*=2.5 and *g*=1 are taken. The following common kernel functions are tested:

- Linear Kernel Function (LKF)

$$K(x_i, x_j) = x_i \cdot x_j \tag{8}$$

- Multinomial Kernel Function (MKF)

$$K(x_i, x_j) = [(x_i \cdot x_j) + 1]^d \tag{9}$$

- Radial Basis Function (RBF)

$$K(x_i, x_j) = \exp(-\frac{\left\| x_i - x_j \right\|^2}{\sigma^2}) \tag{10}$$

We obtain the simulation results of the classification from the same sample data via different kernel functions and found RBF is the best one for education resources classification issue.

V. Experiments

The Support Vector Machine based Education Resources Automatic Classifier has been applied for the classification of education resources in Course Center System of Shanghai Second Polytechnic University.

The training documents are selected from the corpus in our previous research [12] to compare with the experiment

of Vector Space Model (VSM) classification.

The corpus are manully selected from 9 categories, including Mechanical & Electronic Engineering, Electronic & Electrical Engineering, Computer, Economics & Management, Fundmental Science, Foreign Languages, Humanities, Arts and Environmental Engineering.

TABLE I: THE TOTAL NUMBER OF DOCUMENTS USED FOR TRAINING AND TESTING

Data Sets	Training Documents	Testing Documents
Mechanical & Electronic Engineering	43	77
Electronic & Electrical Engineering	37	68
Computer	46	89
Economics & Management	33	77
Fundamental Science	32	87
Foreign Languages	29	82
Humanities	31	85
Arts	35	91
Environmental Engineering	47	79
Total	333	785

TABLE II: THE TOTAL NUMBER OF DOCUMENTS USED FOR TRAINING AND TESTING

Data Sets	Precise	Recall	F1
Mechanical & Electronic Engineering	83.11	79.78	66.31
Electronic & Electrical Engineering	97.33	87.87	88.16
Computer	80.64	93.31	75.24
Economics & Management	92.04	98.82	90.96
Fundamental Science	94.78	79.15	75.02
Foreign Languages	79.12	82.66	65.40
Humanities	95.24	92.00	87.62
Arts	95.61	95.58	91.38
Environmental Engineering	79.93	97.30	77.77
Avg.	88.98	89.61	79.73

The experiment results are the Precise, Recall and F1 value that are common used in evaluation for text classification systems. The equations are:

$$P = \frac{D_{retrieved\&relevant}}{D_{relevant}} \quad (11)$$

$$R = \frac{D_{retrieved\&relevant}}{D_{retrieved}} \quad (12)$$

where $D_{retrieved\&relevant}$ is the number of retrieved and relevant documents, $D_{relevant}$ is the number of all relevant documents and $D_{retrieved}$ the number of all retrieved documents.

After tests in LIBSVM, the results turns out to be the following:

VI. CONCLUSION

In this paper, a Support Vector Machine based Education Resources Automatic Classifier is put forward.

The method and related algorithms consider the particularity of Education Resources, such as specialty domain terms, leveled knowledge, varying much, containing hypertext *et al.*

In the experiment, the results turn out to be acceptable and applicable in educational resources classification. Furthermore, the result is better than VSM based classification method in our previous research.

And, the method can also be applied to deal with other educational raw data with minor changes or optimization.

ACKNOWLEDGMENT

This paper was supported in part by the National Natural Science Foundation of China under Grant No. 61272036, and Key Disciplines of Software Engineering of Shanghai Second Polytechnic University under Grant No. XXKZD1301.

REFERENCES

[1] I. El-Naqa, Y. Yang, M. N Wernick, N. P. Galatsanos, and R. M. Nishikawa, "A support vector machine approach for detection of microcalcifications," *IEEE Trans. on Medical Imaging*, vol. 21, no. 12, 2002.

[2] Y. Artan and X. Huang, "Combining multiple 2ν-SVM classifiers for tissue segmentation," presented at ISBI 2008, 2008.

[3] S. Lucey, "Enforcing non-positive weights for stable support vector tracking," presented at IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[4] S. Ozer, M. A. Haider, D. L. Langer, T. H. van der Kwast, A. J. Evans, M. N. Wernick, J. Trachtenberg, and I. S Yetik, "Prostate cancer localization with multispectral MRI based on relevance vector machines," presented at ISBI 2009, 2009.

[5] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, Chichester, ISBN: 0-471-03003-1, 1998.

[6] C. M. Bishop, Pattern Recognition and Machine Learning, *Springer*, Heidelberg, 2006.

[7] J.-C. Chu, P.-Y. Liu, and W.-L. Wang, "Improved approach to weighting terms in Web Text," *Computer Engineering and Applications*, vol. 43, pp. 192-194, 2007.

[8] S. Marvin and S. Scott, "Feature engineering for text classification," presented at International Conference On Machine Learning, 1999.

[9] H. Kim, and J. Seo, "Cluster-based FAQ retrieval using latent term weights," *Intelligent Systems*, pp. 58-65, April 2008.

[10] L. Song, X.-L. Li, S. Bai, and S. Wang, "An improved approach to weighting terms in text," *Journal of Chinese Information Processing*, vol. 14, no. 6, pp. 8-13, 2000.

[11] Y.-F. Liu, H. Qi, X.-E. Hu, and Z.-Q. Cai, "A modified weight function in latent semantic analysis," *Journal of Chinese Information Processing*, vol. 19, no. 6, pp. 64-69, 2005.

[12] T. Xia, "A vector space model based education resources automatic classifier," *International Conference on Enterprise Systems*, 2014.

Tian Xia was born in April 1979 in Shanghai, China. His major is educational technology specialty and gets his Ph.D. degree from East China Normal University, Shanghai, China in 2007.

He began working for Shanghai Second Polytechnic University from 2007. Currently, he is the director of Digital Media Technology Department. He published articles concerning term weights algorithms improvements, SVM/VSM improvements, etc. His research interests include natural language processing, computing advertisements, big data mining.

Dr. Xia is the director of Institute of Computer Science and Technology of Shanghai Second Polytechnic University.