Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (13): 2568-2573, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.2568.2573
© 2013 Asian Network for Scientific Information

Massive Test Paper Recognition Using SVM and Shallow Parsing

[1, 2]Dongmei Li, [1, 3]Yan Qin, [1]Na Li and [4]Guangxin Wang
[1]School of Information Science and Technology, Beijing Forestry University, Beijing, 100083, China
[2]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China
[3]HSBC Business School, Peking University, ShenZhen, 518055, China
[4]Department of Psychology, Beijing Forestry University, Beijing, 100083, China

Abstract: Importing test paper questions into the database is a key part of initializing the question bank. This thesis proposes a method based on SVM (Support Vector Machine) and shallow parsing to recognize massive test paper and automatically finish the initialization of the question bank. This approach first use SVM to build a hyperplane to separate test paper into two parts, which are the question numbers and the questions. Secondly, automata model based on the principle of shallow parsing is constructed to judge the question numbers which are recognized by SVM and revises the wrong results. Finally, the successful recognized questions are imported into the database automatically after confirming. A large number of experimental results demonstrate that this method does not need any artificial pre-processing work. It can be used to recognize the Word test paper that contains pictures, tables and formulas. The algorithm is proved to be feasible, effective and adaptable and the recognition rate can achieve 100%.

Key words: SVM, SMO, shallow parsing, automaton, test paper recognition

INTRODUCTION

Taking exams is a main way to verify the standard of the examiner. The propositional person collaborate huge test paper resources from the Internet or accumulation. Managing those resources is an onerous work. Question bank system is proposed in order to reduce the manual work. The questions in question bank can be employed to composite test paper automatically according to the multiple criteria of users.

Nowadays, many scholars have done a large amount of researches on test-paper composition, such as abdicative machine learning (Hwang *et al*., 2008), genetic algorithm (Duan *et al*., 2012; Lin *et al*., 2012; Hwang *et al*., 2005; Meng *et al*., 2006) and greedy algorithm (Hwang *et al*., 2006). However, there is only a little research about the initialization of question bank. How to effectively and automatically translate those test papers into question bank and automatically generate test paper by using those question resources is an important project of CAI (Computer Aided Instruction).

Guo and Li (2009) propose paper texts chunking method, in which automata recognition model is constructed, based on predefined recognition rules, like adding some signs of the test text before the automaton recognizing the test. Moreover, Pan and Kang (2011) propose a method that is based on ANTLR. This method treats the test paper as source code and realizes the paper recognition by means of lexical analysis, syntax analysis and semantic analysis. Both of the two methods can recognize some formats of Word test paper correctly. However, when the test paper contains pictures, tables or formulas, those methods are invalid. Moreover, laborious preprocessing is required before recognizing, such as transforming the paper into fixed format. Therefore the two traditional methods mentioned above have limitations when dealing with test paper in different format.

Test paper which only contains text can be seen as shallow parsing in dealing with the problem of nature language. Shallow parsing, also called partial parsing or chunk parsing is completely opposite to full parsing and does not need to find the complete syntactic parse trees. The main task of shallow parsing is to recognize and analyze some simple structures in sentence (Li *et al*., 2003).

Test paper contains many components: texts, figures, tables, formulas and so on. The shallow parsing is not available for those formats. SVM can be used to recognize this kind of test paper because of the characteristic of test paper's graphic structure. SVM can translate the non-linear space problem to linear space problem based on structural risk minimization principle and it is

Corresponding Author: Dongmei Li, School of Information Science and Technology, Beijing Forestry University, Beijing, 100083, China

good for limited samples, non-linear and higher-dimensional problems. Hence, it can be used to recognize test paper.

A method is proposed to recognize test paper by combining the advantages of SVM and shallow parsing. SVM is primarily used to recognize the question numbers and then an automaton model which is built by the principle of shallow parsing analyses corrects the results recognized by SVM. After all the results have been confirmed by the automaton, the right recognized questions are imported into the database automatically. Experimental results show that this method is available in both English and Chinese test paper in any format. The initialization of the question bank can be accomplished automatically and efficiently.

PRINCIPLE AND APPLICATION OF SVM

SVM is first proposed by Cortes and Vapnik (1995) and is widely used in medicine, image recognition, face recognition or e-learning and other areas (Yan *et al*., 2011; Wang and Yang, 2013; Lin *et al*., 2013; Cui and Li, 2011; Hwang *et al*., 2007; Huang *et al*., 2007). The principle of SVM is to build a hyperplane based on the foundation of Vapnik-Chervonenkis dimension and structural risk minimization. It can be used to classify different kinds of training samples by building hyperplanes (Eq. 1) and the purpose is to find the maximum interval hyperplane (Eq. 2):

$$(w.\ x)+b = 0 \tag{1}$$

$$\text{Min} = \frac{1}{2}\| w \|^2, yi(wTxi + b) \geq 1, i = 1, 2, ..., n \tag{2}$$

where, w denotes the normal of hyperplane, x is the training samples, b is the distance between the origin and hyperplane, y is the classification of each training sample.

But Eq. 2 is only applicable to linear problem. Not all the test paper belongs to linear problem and we cannot separate simply the question numbers and the questions simply by a line. In order to deal with general format test paper, the un-linear problem method (Eq. 3) is used to recognize test paper.

$$\text{Min} = (\frac{1}{2}\sum_{i=1}^{1}\sum_{j=1}^{1} y_i y_j (x_i \cdot x_j)\alpha_i\alpha_j - \sum_{j=1}^{1}\alpha_j), \sum_{i=1}^{1} y_i\alpha_i = 0,$$

$$0 \leq \alpha_i \leq C,\ i = 1, 2, ..., l \tag{3}$$

where, α is Lagrange multiplier vector, C is penalty parameter.

Generally speaking, Eq. 3 is not easy to solve. It needs expensive third party quadratic programming tools. Therefore, Sequential Minimal Optimization (SMO) is proposed to reduce the scale of training set-optimization multivariable problem to two-variable problem. So it needs not to solve convex quadratic programming problem in algorithm. At the same time, the number of iteration may increase.

Each iteration chooses and adjusts two components of vector α(α_i and α_j) and the rest of the components of vector α remain constant. Through solving the optimization problem of α_i and α_j, α^*_i, α^*_j are gotten. Then, they are used to update α and b.

Finally, the decision function is as follow:

$$f(x) = \text{sgn}(g(x)) = \text{sgn}(\sum_{i=1}^{1} y_i\alpha_i^*(x_i \cdot x) + b^*) \tag{4}$$

Sometimes, not all the question numbers are on the same line. So, the Eq. 4 is transformed as follow:

$$f(x) = \text{sgn}(g(x)) = \text{sgn}(\sum_{i=1}^{1} y_i\alpha_i^*(x_i \cdot x) + b^* + X),$$

$$0 \leq X \leq b^* \tag{5}$$

TEST PAPER RECOGNITION ALGORITHM BASED ON SVM AND SHALLOW PARSING

Recognition algorithm description: Usually the contents in a test pape can be classify into three parts: title, type of problem and question. The format of test paper can be defined as the following BNF:

< content>::=<title><type of problem><question>
|<type of problem><question>
<title>::=<Chinese>|<letter>|<number>|<singal>
|<table>
<type of problem>::=<number>|<Chinese>
|<letter><signal>
<question>::=<Chinese>|<letter>|<number>
|<signal>|<picture>|<table>|<formula>
<letter>::=A|B|C|…|a|b|c|…
<number>::=1|2|3|…
<signal>::=!|(|)|_|,|…

From the BNF, we can see it is difficult to get all of their vectors, while it is easy to find all the numbers in test paper and get their vectors. Therefore, the SVM only need to recognize those numbers which belong to the question numbers and the numbers in questions. We only deal with the numbers and the first question number's right element

(mark, symbol, Chinese character and letter) for each question pattern in the test paper. The first question number and its right element in each type of problem are set as known training samples, named as K. The rest numbers in each type of problems are set as unknown training samples, named as UK.

All the training samples in SVM are represented as vectors. Therefore, they are defined as sextuple:

$$TN = (x, y, value, sentence, index, digit)$$

where, TN is the set of elements in test paper, x = ([x]1,[x]2) the vector of TN, y = (1, -1) the classification of TN, if TN belongs to the question number, then y = 1, if not, y = -1, sentence the paragraph that TN belongs to, index TN's location in the paragraph, digit TN's digit, if TN is number, then digit equals the number's digit, if not, digit equals 1.

In SMO algorithm, some parameters need to be set (C, tol, Maxpasses). C means penalty parameter, tol tolerance limits and Maxpasses the number of iterations. Every question pattern is set as the unit of recognition. There are two variables in the known set, so Maxpasses sets as 2. The tests found that when C and tol up to a certain value, it has no effect on the recognition result. Therefore, C and tol can be set as 1 and 0.01, respectively.

SVM identifying rule: The identification rules can be defined as follows:

- The location of all the numbers and the first question's right element in the pattern are recorded. Among them, let the first question's decision function f(x) equals 1 and its right element's decision function f(x) equals-1. They compose K. UK is composed by the rest numbers
- Initializing the error term:

$$E_i = \sum_{j=1}^{n} y_i \alpha_i K(x_i, x_j) + b - y_i \quad (6)$$

- Vectors α_1 and α_2 are selected as the adjust points.
- Judging whether α_1 and α_2 satisfy the KKT conditions. If satisfied, it means that α_1 or α_2 needs to update. Assume α_2 needs to update, then:

$$\alpha_2^{new} = \alpha_2^{old} + \frac{y_2(E_1 - E_2)}{\eta}, \quad \eta = K(x_1, x_1) + K(x_2, x_2) - 2K(x_1, x_2) \quad (7)$$

$$\alpha_2^{new} = \begin{cases} H & \alpha_2^{new} \geq H \\ \alpha_2^{new} & L < \alpha_2^{new} < H \\ L & \alpha_2^{new} \leq L \end{cases} \quad (8)$$

among which:

$$\begin{cases} L = \max\{0, \alpha_2^{old} - \alpha_1^{old}\} & y_1 y_2 = -1 \\ L = \max\{0, \alpha_1^{old} + \alpha_2^{old} - C\} & y_1 y_2 = 1 \end{cases}$$

$$\begin{cases} H = \min\{C, C + \alpha_2^{old} - \alpha_1^{old}\} & y_1 y_2 = -1 \\ H = \min\{C, \alpha_1^{old} + \alpha_2^{old}\} & y_1 y_2 = 1 \end{cases}$$

- Change:

$$\alpha_1^{new} = \alpha_1^{old} + y_1 y_2 (\alpha_2^{old} - \alpha_1^{old}) \quad (9)$$

- Using$\alpha^{new}_1$and α^{new}_2 to modify the value of E_i and b
- If the terminating condition reach Maxpasses, the algorithm stops, or jump to (2)
- According to the values of α* and b*, the decision function can be gotten
- Calculating all elements' y through f(x)

Then, we can get the SVM recognition result. All the question numbers in test paper are marked in red color, but SVM also signs some numbers in questions. This is because only the first question number and the elements on its right side are used to be the samples of K and in order to deal the nonstandard test paper, we extend the value of b. The next step is to find those wrong classified numbers and modify them. Shallow parsing is used in the next step to correct the wrong classification.

Shallow parsing identifying rule: Shallow parsing can only recognize some simple structure of a sentence without the need for identification and analysis of the whole sentence. There is no necessary to recognize all the contents in test paper and it is enough to find out the numbers in all question numbers which are recognized in SVM. Therefore, an automaton based on the principle of the shallow parsing can be built to recognize and analyze the result of SVM and deal with the wrong recognition result of SVM.

It is found that the question number is increased in 1 as a unit by analyzing the format of test paper. With a few exceptions, such as manual input errors, normally the right element of question number is the same as each type of problem. In order to make the algorithm suitable for

most test papers, we construct an automata model based on the principle of shallow parsing to revise the wrong results. The automaton handles the nonstandard test papers according to the following rules. (The meaning of symbols is defined in Table 1):

- **Rule 1:** If the right element of the recognition figure equals to P, it means that this figure is QN. The recognition result is right
- **Rule 2:** If this figure does not satisfy the rule1, but the right element satisfies the context (equaling to S+1). It means that this figure might be QN and needs further analysis. However, if the right element does not satisfy the context (equaling to $\overline{S+1}$), it means this figure is NQ and it is unnecessary to process for further handling
- **Rule 3:** If the figure equals to S+1 and its left element equals to T, it means this figure belongs to QN. If the left element of this figure equals to $\overline{T}$, it means this figure belongs to NQ

The recognition process can be described in Fig. 1.

Initial state B: All the QN's which are recognized in SVM will be the input of the automaton. And in this state, automaton will judge whether the QN's right element equals to P. If the QN's right element equals to P, this number's classification is correct and the program jumps to state D, or jump to state A.

Context state A: In this state, the automation will judge whether the QN satisfies S+1. If satisfies, the program jumps to state J, or the QN's classification is wrong, the program jumps to state F.

Table 1: Symbols in automaton

Symbol	Meaning	Symbol	Meaning
QN	Question No.	NQ	No. in question
T	The first question No.'s left element	$\overline{T}$	Not T
P	The first question No.'s right element	$\overline{P}$	Not P
S	The question No. of previous question	$\overline{S}$	Not S

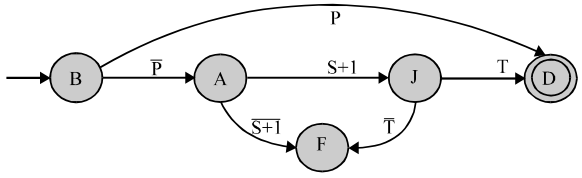

Fig. 1: State diagram of recognition

Modify state F: In this state, the automaton will translate the QN to NQ.

Judging state J: In this state, the automaton will judge if the QN's left elements equal T or not. If not, the QN's classification is wrong, the program jumps to state F, or jump to the state D.

Ending state D: In this state, automaton will finish recognizing the QN or NQ.

The final recognition result judged by the automaton can be seen in Fig. 2. In Fig. 2, question numbers are marked by the red color font and bookmark, questions are marked only by the bookmark. It is clear that all the wrong classified numbers are modified.

EXPERIMENTAL RESULTS

In Eq. 5, the X-value has great impact on the classification result. The bigger the X-value is, the more wrong classification elements will occur. And it will impact the efficiency of the algorithm. But if the X-value is too small, it cannot deal with the unfixed format test paper.

To find the optimal X-value, we take the arithmetic and data structure test papers which are downloaded from the Internet to the test sample. In those test papers, there are 3180 multiple-choices, 3350 blank problems, 2320 true-false, 4500 word problems, 2260 algorithm design problems. The best X-value is evaluated by accurate rate and recall rate. The results are shown in Table 2:

$$P = \frac{\text{Accurate recognition chunking}}{\text{Recognition chunking}} \times 100\% \quad (10)$$

$$R = \frac{\text{Accurate recognition chunking}}{\text{Real chunking}} \times 100\% \quad (11)$$

From the Table 2, the minimum misclassification rate is obtained when X-value is b/1.5. Whatever the X-value is given, the misclassification rate of the automaton is zero. The final recognition accuracy rate can achieve 100%.

To estimate the recognition efficiency and accuracy rate of the algorithm, numerous test papers with pictures,

Table 2: Accurate rate and the recall rate of recognition result

X	P1 (%)	R1 (%)	P2 (%)	R2 (%)
0	45.52	100	100	100
b/(1.5)	56.06	100	100	100
b/2	42.77	100	100	100
b/3	38.19	100	100	100
b/4	36.37	100	100	100
b/5	35.07	100	100	100

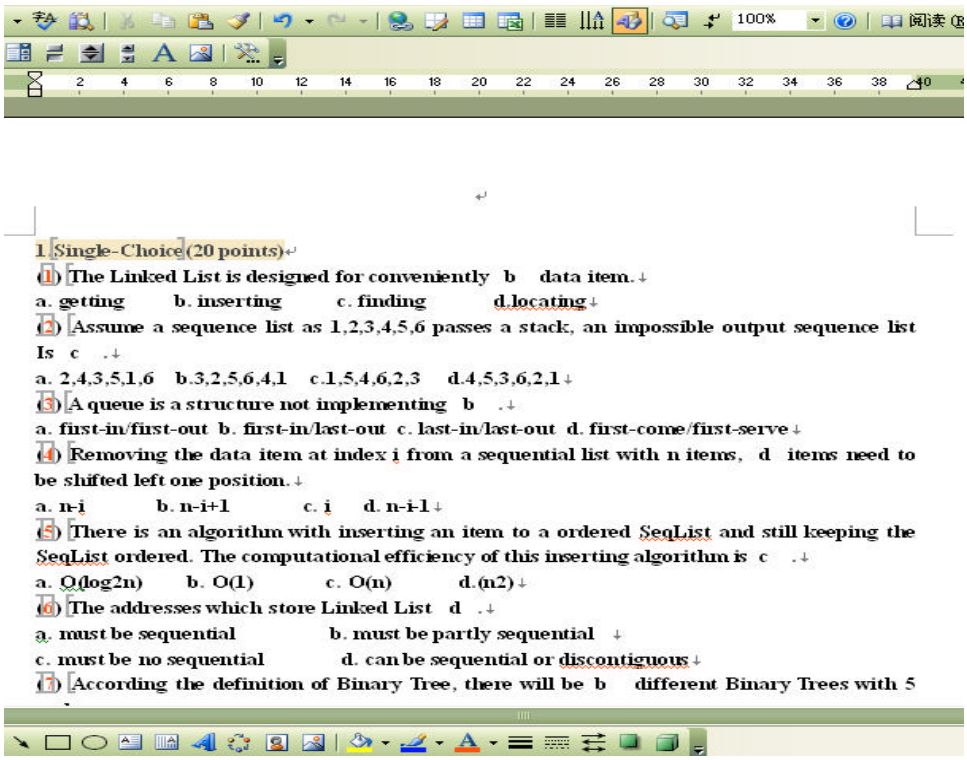

Fig. 2: Result of automaton recognition

formulas and tables are recognized. It is obvious that our approach can accurately recognize test paper which contains pictures, formulas and tables.

CONCLUSION

A new method, which is based on the principle of SVM and shallow parsing, is proposed to recognize massive Word test paper. It consists of two steps, firstly, SVM classifiers are used to classify the question numbers and the numbers in questions. Secondly, shallow parsing is used to judge the recognition result and correct wrong classifications. The final results validate the feasibility of the proposed method. However, the algorithm needs automaton to deal with the preliminary results recognized by SVM. For future work, we intend to improve the performance of SVM classification and reduce the work of automaton so that the efficiency of the algorithm will be further improved.

ACKNOWLEDGMENT

This study is supported by the Education of Humanities and Social Science Research on Youth Fund Project (11YJC190024) and the National Nature Science Foundation of China (No. 61170628).

REFERENCES

Cortes, C. and V. Vapnik, 1995. Support-vector networks. Machine Learn., 20: 273-297.

Cui, G.Q. and J.T. Li, 2011. Face recognition using support vector machines. Comput. Sci., 30: 67-70.

Duan, H., W. Zhao, G.G. Wang and X.H. Feng, 2012. Test-sheet composition using analytic hierarchy process and hybrid metaheuristic algorithm TS/BBO. Math. Prob. Eng., Vol. 2012 10.1155/2012/712752

Guo, K.H. and L.W. Li, 2009. Study of massive paper texts chunking based on rules. Appl. Res. Comput., 26: 1391-1393.

Huang, C.J., S.S. Chu and C.T. Guan, 2007. Implementation and performance evaluation of parameter improvement mechanisms for intelligent e-learning systems. Comput. Educ., 49: 597-614.

Hwang, G.J., B.M.T. Lin, H.H. Tseng and T.L. Lin, 2005. On the development of a computer-assisted testing system with genetic test sheet-generating approach. IEEE Trans. Syst. Man Cybernetics C, 35: 590-594.

Hwang, G J., B.M.T. Lin and T.L. Lin, 2006. An effective approach for test-sheet composition with large-scale item banks. Comput. Educ., 46: 122-139.

Hwang, G.J., P.Y. Yin, Y.T. Wang, J.C.R. Tseng and G.H. Hwang, 2007. An enhanced genetic approach to optimizing auto-reply accuracy of an e-learning system. Comput. Educ., 51: 337-353.

Hwang, G.J., H.C. Chu, P.Y. Yin and J.Y. Lin, 2008. An innovative parallel test sheet composition approach to meet multiple assessment criteria for national tests. Comput. Educ., 51: 1058-1072.

Li, S.J., Q. Liu and Z.F. Yang, 2003. Chunk parsing with maximum entropy principle. Chinese J. Comput., 26: 1734-1738.

Lin, H.Y., J.M. Su and S.S. Tseng, 2012. An adaptive test sheet generation mechanism using genetic algorithm. Math. Prob. Eng., Vol. 2012 10.1155/2012/820190

Lin, C., X. Qin, G.L. Zhu, J.H. Wei and C. Lin, 2013. Face detection algorithm based on multi-orientation gabor filters and feature fusion. Telkomnika Indonesian J. Electr. Eng., 11: 5986-5994.

Meng, A., L. Ye, D. Roy and P. Padilla, 2006. Genetic algorithm based multi-agent system applied to test generation. Comput. Educ., 49: 1205-1233.

Pan, X. and M.N. Kang, 2011. Research on examination paper recognizing and importing system based on ANTLR. Electron. Des. Eng., 19: 45-49.

Wang, Y.L. and H. Yang, 2013. Machine learning to design full-reference image quality assessment algorithm. Telkomnika Indonesian J. Electr. Eng., 11: 3414-3421.

Yan, J., J. Li and X. Gao, 2011. Chinese text location under complex background using gabor filter and SVM. Neurocomputing, 74: 2998-3008.