Scientific Research Publishing

Journal of Computer and Communications, 2017, 5, 1-13
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Mathematical Expression Extraction in Text Fields of Documents Based on HMM

Xuedong Tian*, Ruihan Bai, Fang Yang, Jinyuan Bai, Xinfu Li

School of Cyber Security and Computer, Hebei University, Baoding, China
Email: *Xuedong_tian@126.com

How to cite this paper: Tian, X.D., Bai, R.H., Yang, F., Bai, J.Y. and Li, X.F. (2017) Mathematical Expression Extraction in Text Fields of Documents Based on HMM. *Journal of Computer and Communications*, **5**, 1-13.
https://doi.org/10.4236/jcc.2017.514001

Received: November 7, 2017
Accepted: December 9, 2017
Published: December 12, 2017

Abstract

Aiming at the problem that the mathematical expressions in unstructured text fields of documents are hard to be extracted automatically, rapidly and effectively, a method based on Hidden Markov Model (HMM) is proposed. Firstly, this method trained the HMM model through employing the symbol combination features of mathematical expressions. Then, some preprocessing works such as removing labels and filtering words were carried out. Finally, the preprocessed text was converted into an observation sequence as the input of the HMM model to determine which is the mathematical expression and extracts it. The experimental results show that the proposed method can effectively extract the mathematical expressions from the text fields of documents, and also has the relatively high accuracy rate and recall rate.

Keywords

Mathematical Expression Extraction, Hidden Markov Model, Text Fields, Documents, Symbol Combination Features

1. Introduction

Extracting the information we need from the text in documents and converting them into structured data for storing in database is the premise of their further utilization [1]. At present, there are mainly three kinds of text information extraction technology: dictionary based model [2], rule based model [3] and statistics based model [4].

Hidden Markov Model [5] is one of the statistical models which obtains its model parameters through training instead of prior structural dictionaries and rules. It is established easily with stronger applicability and higher accuracy rate and recall rate. Therefore, it has got more attention of researchers, and widely been used in word segmentation [6], speech recognition [7], information extraction [8] and other fields.

Yu *et al.* [9] proposed an algorithm based on Hidden Markov Model for extracting the information of the header and reference of scientific papers in Chinese. At first, the algorithm segmented the paper header into semantic blocks by utilizing semicolon, enter, comma and other punctuation marks, and then extracted information from the semantic blocks. In order to make HMM more practical, the algorithm used the method of vertical merger and horizontal merger to combine HMM states when the model structure was being trained. In this way, an optimized HMM was obtained which had effectively improved the extraction accuracy rate and recall rate.

Liu [10] proposed an extraction method that can extract text information from different fields and data sets based on HMM. Firstly, the initial probability and transition probability parameters were divided into several templates for training. Based on this, a text information extraction algorithm based on multi-template HMM was proposed. It can effectively solve the problem that the training data was diversified and the model parameters were hard to learn, and has higher accuracy rate and recall rate. In addition, a new algorithm of text information extraction based on maximum entropy HMM was proposed through combining rule based method and statistic based method according to the information of context features and words meaning, which obtained a better extraction result.

Liu [11] proposed an information extraction method based on maximum entropy HMM (ME-HMM) and ontology technology for accurately extracting useful information from comprehensive evaluation of books. ME-HMM was used to some pre-processing works, such as nominal marking and named entity recognition. The information extraction model based on ontology technology was used for the extraction of professional sentence and addresses to solve the problems of information association and unclear underlying meaning. The experimental results show that this method can improve the accuracy rate of information extraction.

Du *et al.* [12] put forward a method of text information extraction based on mixed HMM to get the place names from the metadata of documents. This method considered the dependence of transition probability at a moment on previous and latter states. Two pass decoding of positive sequence and negative sequence were performed by utilizing Viterbi algorithm as decoding. The experimental results show that the method can achieve a better information extraction result.

Shuang and Sun [13] proposed an improved information extraction method based on HMM. The transition probability and output probability of the improved model were not only dependent on the current state, but also can be corrected according to the forward and backward dependency which can further improve the information extraction quality. The experimental results show that the proposed method can further improve the extraction quality.

In a word, the HMM has been widely used in text information extraction and played a tremendous role. However, there are less researches employing HMM on the mathematical expression extraction from text fields.

Because of the complex syntax rules of the mathematical expressions, if we use the rule based method of text field extraction to extract mathematical expressions, complicated rules should be made and the applicability of this method is limited. Based on the above considerations, this paper proposes a method to extract mathematical expressions from the unstructured text fields of documents based on HMM which could effectively avoid these problems. Firstly, the HMM model is constructed according to the symbol combination feature of the mathematical expression. Then, the extracted text is pre-processed. Finally, the trained model is used to extract the mathematical expressions automatically. Simulation results show that this method can automatically, rapidly and effectively extract mathematical expressions from the text fields.

2. The Description of the Problem

It is found that the structured web pages express formulae through description language such as LaTeX [14], MathML [15], OpenMath [16] and so on which generally have fixed embedding modes and separating labels. For example, in LaTeX documents, the embedded formulae are labeled as "$…$", "\[…\]", "\begin{math}…\end{math}". While the isolated formulae are labeled as "$$…$$", "\[…\]", "\begin{align}…\end{align}", etc. Therefore, the mathematical expressions in the structured documents can be extracted according to the labels. Cui *et al.* [17] summarized the characteristics of various kinds of formats of formulae exist on web pages. And put forward a method of mathematical formula extraction based on heuristic rules. Chen *et al.* [18] proposed a method to automatically analyze and extract the features of LaTeX expressions and filtering formulae according to their features. In this method, the extraction accuracy was improved efficiently.

Different from structured text like web pages, the mathematical expressions existing in unstructured text have no clear labels. We cannot extract mathematical expressions from these kinds of documents directly with labels. The content of text consists of characters, numbers, punctuations and other printable symbols. And the combination modes of these symbols are many and varied. There are also no observable rules to separate mathematical expressions from ordinary text. In addition, the sources of text are varied, the typesetting forms and expression styles of text are also different, and the content of text is very large sometime. All of these reasons make great challenges and difficulties to extract mathematical expressions from ordinary text. So it is necessary to research the method for extracting mathematical expressions from ordinary text without special labels.

3. Algorithm for Extracting Mathematical Expressions in Text Field of Documents Based on HMM

3.1. The Overall Framework of the Algorithm

According to the characteristics of mathematical expressions in the text fields of documents, this paper proposes an automatic extraction algorithm of the mathematical expressions based on HMM. The overall framework is shown in **Figure 1**.

The process of extracting mathematical expressions in text fields of documents based on HMM can be divided into two parts described as follows:

1) Training model

Input: Training samples (the mathematical expressions).

Output: HMM model.

Step 1: Get the training samples, that is, the correct mathematical expressions.

Step 2: Set up the HMM status value and observation value for expressions.

Step 3: Use Baum-Welch algorithm [19] to learn the parameters for training data and set up the model.

Step 4: When the training generation number meets the preset value, stop training; otherwise, go to Step 3.

2) Testing model

Input: Trained HMM model and test samples (the text which contains mathematical expressions to be extracted).

Output: Mathematical expressions.

Step 1: Obtain the text that needs to be processed.

Step 2: Text pre-processing.

Step 3: Extract mathematical expressions.

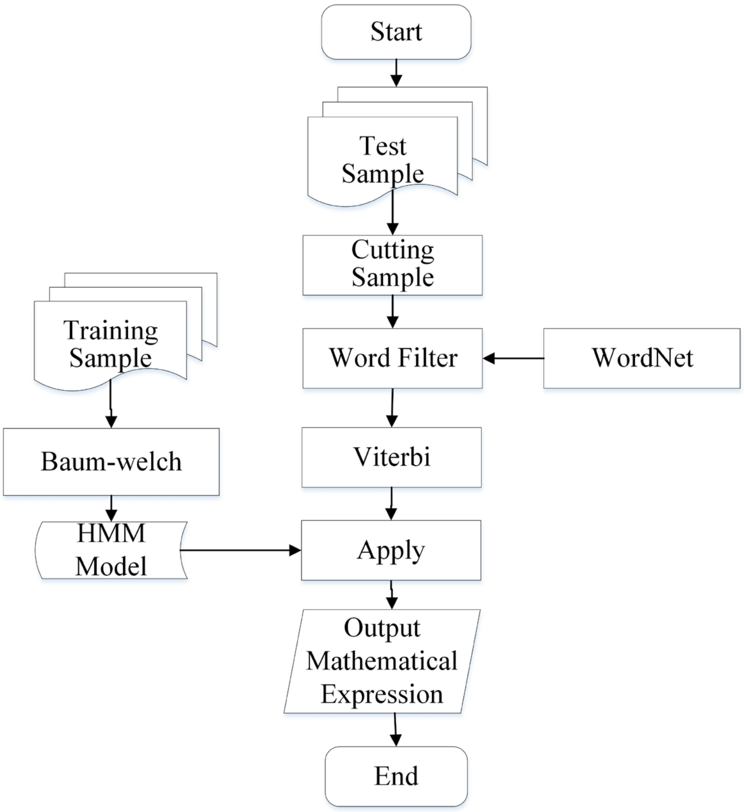


Figure 1. Overall framework of proposed method.

3.2. Construction of HMM Model and Its Parameter Learning

Some symbols of the mathematical expressions in text represented in Unicode codes are shown in Table 1.

Hidden Markov Model (HMM) [20] is the model of a double random process based on statistics. To construct the HMM model, we should first determine the structure of the model, that is, the number of contained states and the conversion between each state. According to the Table 1, we define a total of eight states of HMM as: the variables (numbers, letters, Greek letters), monocular operation (the operator has one operand only), binocular vision (the operator has two operations), punctuation, left and right delimitations, delimiter in a set, superscripts and subscripts, denoted as $S=\{E,U-O,B-O,P,G_1,G_2,B-D,C,Q\}$. The observed values of the HMM model are the mathematical symbols shown in the mathematical expressions. This paper defines that the observation values corresponding to each state which are shown in Table 2.

In addition to the above observations, we define the trigger word dictionary as follows:

Definition 1: Trigger_ W, the trigger word, is an important part of a specific event that can describe its special attribute information.

In mathematical expression extraction, Trigger_ W = {sin, cos, Tan, cot, sec, csc, arcsin, arccos, arctan, arccot, arcsec, arccsc}. Trigger_ W is employed as the

Table 1. Special symbols of mathematical expressions.

Symbolic	Meaning	Symbolic	Meaning
+, −, *, /, ×, ÷	Four arithmetic symbols	(), [], {}	Bracket
=, ≠, ≤, ≥, ≈, <, >	Equivalent relational symbols	∫, ∮	Integral symbol
∩ , ∪	Set symbols	||	Absolute value sign
%	Percent symbol	!	Factorial symbol
Σ, Π	Summation symbol and quadrature symbol	^, _	Square, Subscript
:	Division symbol	∃	Existential symbol
√	Radical sign	&	And

Table 2. Observation values corresponding to each state.

State	Observation values
E	Numbers, a-z, A-Z, Greek letters, begin with digitals and end with letters, begin and end with letters
$U-O$	+, −, !, %, &
$B-O$	+, −, =, /, :, *, <, >
P	!, :, ;
G_1	{, [, (, |
G_2	},],), |
$B-D$	, , ;
Q	^, _

prior knowledge in knowledge base for trigonometric function, inverse trigonometric function extraction.

Taking the formula "$a!+4b\times(c\div 5)=-d$" as an example, the corresponding state value and the observed value are shown in **Figure 2**.

When the HMM model is constructed, it is assumed that there are a meaningless start state (start) and a meaningless end state (end) as the preparatory state and final state of the model. Taking the mathematical expressions "$a!+4b\times c=-d$", "$(a-b)\div 5d=-c$", "$3a+4b<(-ab)\times c$" as an example, it's HMM model structure is shown in **Figure 3**.

After determining the model structure, we need to find the transition probability of state and the output probability of observation. In this experiment, the Baum-Welch algorithm was used to calculate the model parameters, and the specific steps are as follows:

Step 1: Define two intermediate variables: $\xi_t(i,j)$ and $\gamma_t(i)$. Where $\xi_t(i,j)$ is the probability of the state i and the state j corresponding to the model $\lambda=\{A,B,\Pi\}$ and the observed sequence $O=\{o_1,o_2,\cdots,o_T\}$ at the moment t and the moment $t+1$. $\gamma_t(i)$ is the state probability at the moment t of the

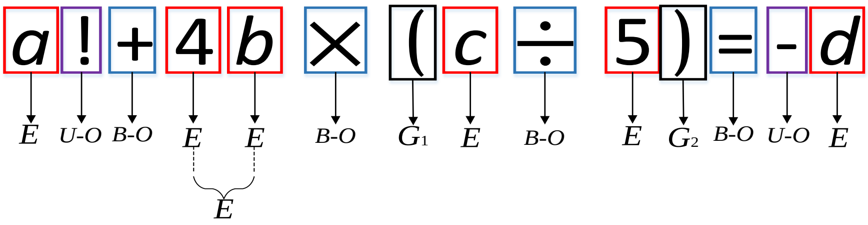


Figure 2. State value and observed value of an expression.

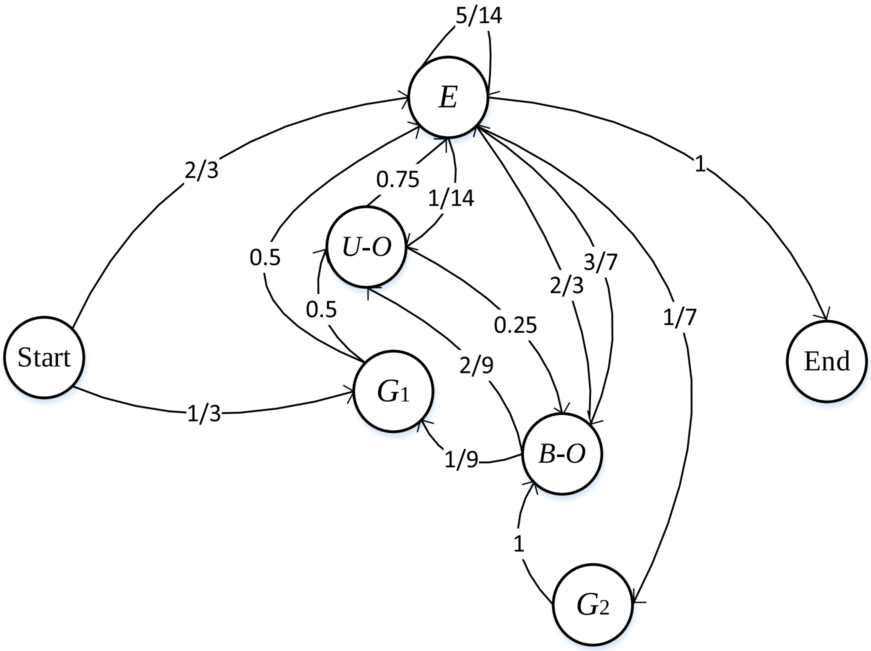


Figure 3. Instance diagram of state value observation value correspondence.

corresponding state i in the case of given model and observation sequence [13]:

$$\xi_t(i,j)=\frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)} \tag{1}$$

$$\gamma_t(i)=\sum_{j=1}^{N}\xi_t(i,j)=\frac{\alpha_t(i)\beta_t(i)}{P(o|\lambda)}=\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)} \tag{2}$$

Step 2: Initialize [10]: $\overline{\pi}_i=\gamma_1(i)$, the expected value of the state S_i at time $t=1$ is $\lambda=\{A_0,B_0,\pi\}$.

Step 3: Iterative compute: Let $\lambda_0=\lambda$

$$\overline{\pi}_i=\gamma_1(i),1\le i\le N \tag{3}$$

$$\overline{a}_{ij}=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{4}$$

$$\overline{b}_j(k)=\frac{\sum_{t=1}^{T}\gamma_t(j)\times\delta(o_t,v_k)}{\sum_{t=1}^{T}\gamma_t(j)} \tag{5}$$

where,

$$\begin{cases}\delta(o_t,v_k)=1,o_t=v_k\\ \delta(o_t,v_k)=0,o_t\ne v_k\end{cases} \tag{6}$$

Step 4: The initial value is substituted into the above equations, and the new parameters are calculated. The iteration is repeated until the parameter converges, the method of judging convergence is:

$$\left|\log\left(P(o|\lambda_{i+1})\right)-\log\left(P(o|\lambda_i)\right)\right|<\varepsilon \tag{7}$$

where the threshold ε is generally set to $1e-6$.

3.3. Extraction of Mathematical Expressions

Extracting mathematical expressions from text fields of documents can be implemented with the following steps:

Step 1: Design a crawler, and the depth of crawler is 20; select website home pages which are relevant to mathematics as the crawler entrance; start the crawler, and save the acquired pages to a local file.

Step 2: For the web pages crawled from the Internet, block the web pages at first [21]. Remain the main part and delete the navigation blocks and link blocks. As there are a lot of HTML labels in web pages, the blocked text will be cleaned through removing the useless noise data in order to reduce the noise in samples and improve the accuracy rate of mathematical expression extraction, which is shown in **Table 3**.

Table 3. Noise types and processing methods.

Noise types	Processing methods
<sup></sup>; <sub></sub>	replaced with “^”; “_”
mathematical expressions with label <math></math>	regular matching filtering <math[\\s\\S]{1,}?</math>
<p></p> label	regular matching filtering: <.*?>
&nbsp; &amp; &gt; &lt	replaced with “blank”; “&”; “>”; “<”

Step 3: Segment the text which have been removed the useless noise data. After segmentation, use the word library of WordNet [22] (containing about 350 thousand words) to filter the text. As shown in **Figure 4**, the filtered text will be taken as the input of HMM model.

Step 4: Use the HMM model which has been constructed and the Viterbi algorithm to determine whether the expression in text is a mathematical expression or not, if so, the expression will be extracted.

For example, the proposed method analyzes the sentence “There’s good reason why mathematics formulas are not patentable! It’s really easy {4 ghs/} to come up with new formulas. Like $a+b=c+d$. I bet you won’t find that in any textbook, because 56 + 4 = 60 it’s not something really important.” through the following steps:

Step 1: Initialize: Build training samples and test samples.

Step 2: Train HMM model by training samples.

Step 3: Confirm the model structure, and evaluate parameters by Baum-Welch algorithm.

Step 4: Do some pre-processing works for the text.

```
String content = mwContent,html().replaceAll(“noise”, “replace word”)
//remove the useless noise from the text
String [] tempArr = sb.toString ().replaceAll (“[  ]{2,}”, “  ”).replaceAll
(“\\[[0-9a-zA-Z]{1,}]\\”, “ ”).split(“ ”);
//cut the text according to the blank
For each vocabulary in tempArr {
    If (wordSet contains vocabulary) {continue}
    else {save vocabulary until the next math}
    }
//Use the WordNet word library to filter the text, and then save the filtered
text in fmList.
```

Step 5: Take the fmList as the input for HMM model.

Step 6: Use the Viterbi algorithm to determine whether the expression is a mathematical expression.

Step 7: If so, the expressions will be extracted.

4. Experiment Results and Analysis

In order to prove the availability of proposed method based on HMM model in

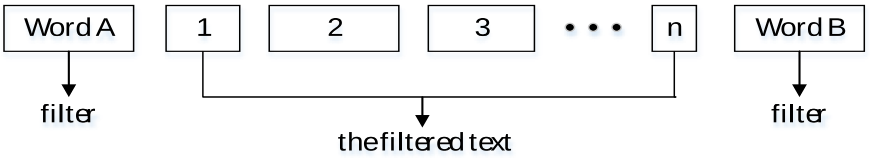


Figure 4. Word filtering.

this paper, the simulation experiment is carried out according to the above steps with the development environment of the open source Jahmm-0.6.1 framework and MyEclipse software.

The training samples in this simulation experiment are 13,423 mathematical expressions from mathematical textbooks of elementary school. The testing samples are 2100 web pages from the Internet. Each web page is taken as a sample. According to the construction rule of HMM model described above, the new model parameters can be obtained as $\lambda = \{A, B, \Pi\}$. Part of the observation probability distribution matrix after training is shown in Figure 5.

The experimental results will be evaluated by two evaluation indexes: accuracy rate and recall rate. The calculation methods of two evaluation indexes are expressed as follows:

$$\text{Accuracy rate} = N_{\text{extracted}} / N_{\text{marked}} \tag{8}$$

$$\text{Re call rate} = N_{\text{extracted}} / N_{\text{all}} \tag{9}$$

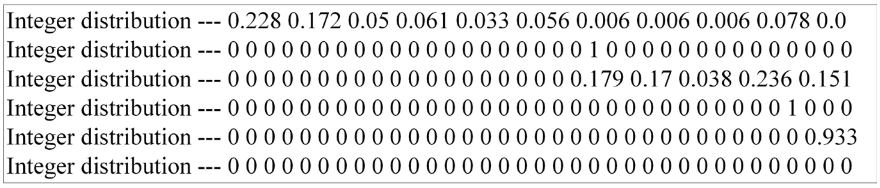
where $N_{\text{extracted}}$ is the number of the mathematical expressions extracted correctly; N_{marked} is the number of the marked mathematical expressions in the text; N_{all} is the number of all mathematical expressions in the text.

As shown in Table 4, the accuracy rate and recall rate of the proposed algorithm have some slight oscillations in the extraction process, but the whole situation tends to be stable. With the increase of text number, the accuracy rate and recall rate will also increase and tend to be stable finally, but they may decrease when the text number is too large. Because of the limited training samples and various mathematical expressions, it is very difficult to describe all cases by only a model, so the extraction model should be expanded further.

Besides, by experimental analysis, the influence factors are those texts which are similar to mathematical expressions, such as time symbol (2007-03-02), name symbol (Jason-Leon), number symbol (N68-32), unregistered words symbol (e.g. e-mail), etc.

For verifying the performance of our method, we built a simulation system which employed the algorithm proposed in Reference [18]. It is worth nothing that due to the differences of experimental environment, experimental data such as the mathematical expression format, and the limitations of our comprehensions on the original paper, the comparison results could not completely meet the results of the original paper and could only roughly reflect the situations of the method.

The comparison results about accuracy rate and recall rate are shown in Figure 6.



```
Integer distribution --- 0.228 0.172 0.05 0.061 0.033 0.056 0.006 0.006 0.006 0.078 0.0
Integer distribution --- 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Integer distribution --- 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.179 0.17 0.038 0.236 0.151
Integer distribution --- 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
Integer distribution --- 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.933
Integer distribution --- 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Figure 5. Probability distribution matrix of training results.

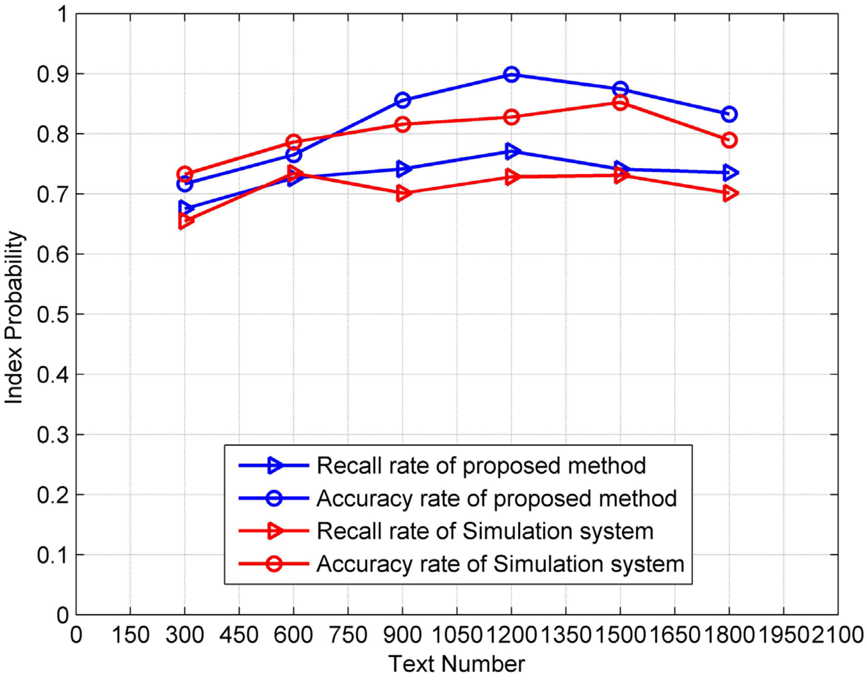


Figure 6. Comparison results on the accuracy rate and recall rate.

Table 4. Extraction results of mathematical expressions.

Methods		300	600	900	1200	1500	1800
Proposed method	Recall rate	0.6753	0.7264	0.7415	0.7709	0.7410	0.7354
	Accuracy rate	0.7167	0.7650	0.8557	0.8986	0.8741	0.8325
Simulation system	Recall rate	0.6549	0.7344	0.7011	0.7282	0.7310	0.7013
	Accuracy rate	0.7328	0.7862	0.8157	0.8278	0.8523	0.7893

It can be seen from Figure 6 that the proposed method has better recall rate and accuracy rate. Moreover, with the increase of text number, the advantage of the proposed method is more obvious. The reason is that the rule based method need to locate the symbols of mathematical expressions and scan back and forth to extract mathematical expressions. When it identifies the scope of texts, the error of taking ordinary text as mathematical expressions would easily occur. In addition, with the increase of the number of mathematical expressions, the rule based method need to develop a large number of combination rules. Once the combination rules are incomplete, the recall rate will be affected, while the proposed method could effectively avoid the problems.

For verifying the time efficiency of proposed method, different numbers of mathematical expressions were selected to test the extraction time. The test data is shown in **Figure 7**.

From **Figure 7**, we can see that the average extraction time is 0.0011 ms when the number of mathematical expressions in text equals 5 and 0.0053 ms when the number of mathematical expressions in text equals 45 respectively. While the average extraction time of the simulation system is longer than the proposed method. The reason is that the compared method need to locate the mathematical expressions and scan back and forth firstly, and carry out many filtering operations before extraction. The proposed method can avoid these processes and improve the extraction efficiency obviously. These two curves have some overlapping data at the first phase. The reason is that when the number of mathematical expressions is relatively great, the advantage of proposed method is obvious, and the extraction time is less than simulation system. In contrast, the extraction time is about the same when the number of mathematical expressions is small.

5. Conclusion

To solve the problem that the mathematical expressions in unstructured text fields of documents is hard to be extracted automatically, rapidly and effectively, a novel method based on HMM was proposed. Through considering the characteristics of mathematical expressions, this method can realize the extraction of the mathematical expressions in text fields of documents with relative high accuracy rate and recall rate. However, there are some shortcomings of the

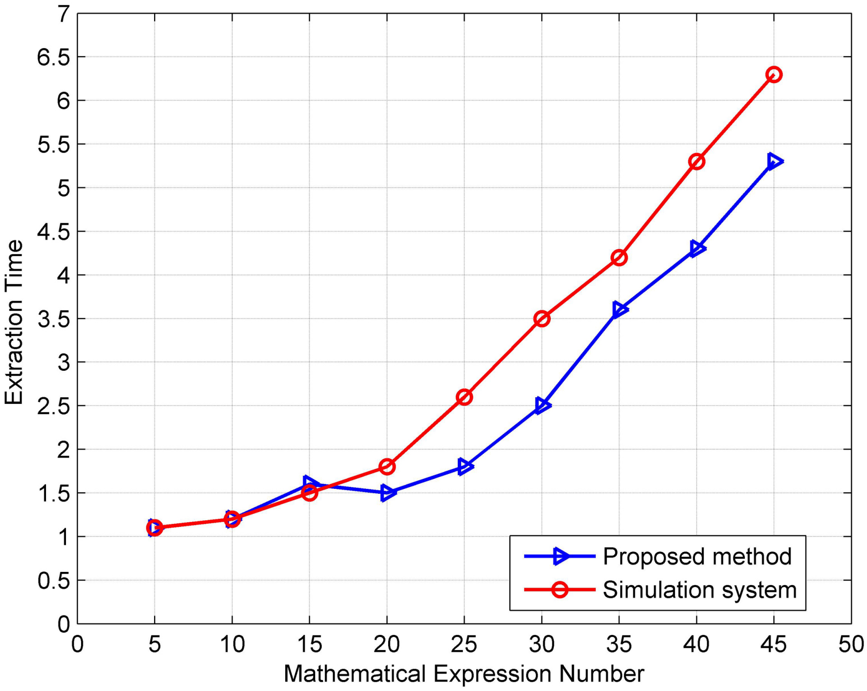


Figure 7. Comparison result on the extraction time.

mathematical expression extraction method based on HMM. Firstly, this method needs plenty of training samples and takes a lot of training time. Secondly, if the states of HMM is not comprehensive enough, the extracted result will be not good. We shall further improve the observed values and state values of HMM model and optimize the model parameters to make the model be more suitable for more mathematical expressions and achieve better extraction results.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 61375075). The Key Project of the Science and Technology Research Program in University of Hebei Province of China (Grant No. ZD2017208; ZD2017209).

References

[1] Guo, X.Y. and He, T.T. (2015) Survey about Research on Information Extraction. *Computer Science*, **42**, 14-17.

[2] Zhu, H.H. and Yu, Q.S. (2014) Study on the Extraction of Chinese Microblog Subjective Sentences Based on Lexicon and Corpus. *Journal of East China Normal University* (*Natural Science*), No. 4, 62-68.

[3] Yu, C., Mao, Z. and Gao, S. (2017) An Approach of Extracting Information for Maritime Unstructured Text Based on Rules. *Traffic information and safety*, **35**, 40-47.

[4] Li, Q. and Chen, Y.P. (2010) Personalized Text Snippet Extraction Using Statistical Language Models. *Pattern Recognition*, **43**, 378-386. https://doi.org/10.1016/j.patcog.2009.06.003

[5] Zhou, C. and Li, S. (2011) Research of Information Extraction Algorithm based on Hidden Markov Model. 2010 2*nd International Conference on Information Science and Engineering* (*ICISE*), Hangzhou, 4-6 December 2010, 1-4.

[6] Wang, Q.F. (2016) Research on Chinese Word Segmentation Based on Hidden Markov Model. *Wireless Internet Technology*, **13**, 106-107.

[7] Yuan, L.C. (2008) A Speech Recognition Method Based on Improved Hidden Markov Model. *Journal of Central South University* (*Science and Technology*), **39**, 1303-1308.

[8] Zhu, W.H., Lu, Y. and Liu, B.B. (2010) Improvement of Web Information Extraction Algorithm Based on HMM. *Computer Science*, **37**, 203-206.

[9] Yu, J.D., Fan, X.Z. and Yin, J.H., *et al.* (2007) Information Extraction from Chinese Research Papers Based on Hidden Markov Model. *Computer Engineering*, **33**, 190-192.

[10] Liu, Y.Z. (2003) Algorithm Research for Text Information Extraction Based on Hidden Markov Model. Master's Thesis, Hunan University, Changsha.

[11] Liu, Z. (2016) Research on the Text Information Extraction of Author Relevant Information. Master's Thesis, Northeast Normal University, Nanjing.

[12] Du, Q.X., Wang, H.G. and Shao, Z.Z., *et al.* (2017) Place Name Extraction Method of Literature Metadata Based on the Hybrid HMM. *Computer & Digital Engineering*, **45**, 101-106.

[13] Shuang, Z. and Sun, L. (2017) Research and Application for Web Information Extraction Based on Improved Hidden Markov Model. *Computer Applications and*

Software, **34**, 42-47.

[14] Gurari, E.M (2004) TEX4ht: HTML Production http://www.tug.org/TUGboat/tb25-1/gurari.pdf

[15] Wang, Q. and Liang, J. (2004) Mathematical Markup Language. Modern Educational Technology, **14**, 63-66.

[16] Strotmann, A. and Kohout, L. (2000) Openmath. *ACM SIGSAM Bulletin*, **34**, 66-72. https://doi.org/10.1145/362001.362024

[17] Cui, L.W., Su, W., Guo, W., *et al*. (2011) Extraction of Web Mathematical Formulas Based on Nutch. *Journal of Guangxi Normal University*: *Natural Science Edition*, **29**, 167-172.

[18] Chen, L.H., Su, W., Cai, C., *et al*. (2014) Research of Extraction Method of Web Mathematical Formula Based on LaTex. *Computer Science*, **41**, 148-154.

[19] Liu, B.B. (2008) Research and Improvement of Web Information Extraction Method Based on HMM Model. Chongqing University, Chongqing.

[20] Zhang, Y. (2014) The Algorithm Research of Chinese Information Extraction Based on the Hidden Markov Model. University of Science and Technology Liaoning, Anshan.

[21] Ren, L.F. (2012) Design and Implementation of Education News Webpage Information Extraction System. South China University of Technology, Guangzhou.

[22] Geum, Y. and Park, Y. (2016) How to Generate Creative Ideas for Innovation: A Hybrid Approach of WordNet and Morphological Analysis. *Technological Forecasting & Social Change*, **111**, 176-187. https://doi.org/10.1016/j.techfore.2016.06.026