Metadata Extraction Approach of PDF Documents Based on Measurement Fusion

Junmin Zhao
Henan University of Urban Construction/Institute of Computer Science and Engineering, Pingdingshan, China
Email: zjm20092@yeah.net

Huazhong Liu
Beijing Mapbar Technology Co. Ltd. Beijing, China

***Abstract*—To deal with the problems of low precision rate and weak adaptability in the existing metadata extraction methods, a novel metadata extraction approach is proposed based on measurement fusion rule in this paper. First, the features of the document header are extracted, the three statistical learning methods such as HMM, SVM and CRF are respectively employed to train the labeled data set, and corresponding metadata extraction models are constructed. Then, the results from three extraction models are fused by the sum rule so as to achieve the accurate metadata extraction of documents. Finally, we dynamically update the three extraction models to guarantee the effectiveness of the ensemble models by the time period statistics-based method. Experiments on different datasets are conducted and the comparative results of these extraction methods are presented; Experimental results show that the proposed approach not only improves the precision of metadata extraction, but also enhances the adaptability.**

***Index Terms*—Metadata Extraction; Statistical Learning; Measurement Fusion; Posterior Probability; Sum Rule**

I. INTRODUCTION

When the related administrators create digital resources database by using the Open Access (OA) journal articles as the information source, how to quickly extract the metadata of the document with high quality is the key issue for automatically generating digital database. Generally, the metadata of an article includes title, author, abstract and key words. By utilizing metadata to organize and manage the OA journals documents in digital resources database, the precision and efficiency of document retrieval can be improved.

Automatic metadata extraction is a popular research topic in the field of library digital resource construction. There are two major realization methods such as rule-based method and machine learning-based method. For the former, the rule set must be constructed and extracted in advance, and then use the rules to extract metadata from the documents. For instance, the CiteSeer system [1] and [2] made use of this method to extract metadata from PDF documents. But this method needs some artificial processes to extract rules and requires the rule-makers owning good knowledge of the application fields. Besides, the rules will be incompatible if extraction targets change. On the other hand, the machine learning-based method train the learning models by a large number of training data, and the trained model can automatically deal with the new documents. For example, Seymore [3] and Liu et al. [4] both proposed some metadata extraction methods based on Hidden Markov Model (HMM), but they did not consider the relevance between state transition probability, the output probability of observed value and the state of historical data in the model. To solve the problem, Ojokoh et al. [5] proposed a method that promoted extraction performance based on third order HMM. Zhou et al. [6] extracted metadata of documents by using the maximum entropy Markov model, which integrates context information and information contained in the words. This method improves the extraction precision, but brings label bias problem. Peng et al. [7] proposed a method based on conditional random field, which can effectively use context features and solve the label bias problem. It can improve the precision of extraction, but it can cause the missing of some low frequency words and inaccurate identifying of high frequency words. Lin et al. [8] extracted metadata from academic articles in clinical medicine by conditional random field. The extraction effect was good for text including specific parameters, but it was not so effective for extracting the author's specific information. Han et al. [9] classified the document blocks by using support vector machine (SVM) method and regarded each metadata as one class. This method can effectively extract document metadata, but can only deal with small samples. Marinai et al. [10] extracted document metadata by using the neural network classifier. Due to the limitation of conversion tool, he only extracted some information about authors and titles from the recent conference articles. Zhang [11]. proposed a hybrid metadata extraction model (SVM+BiHMM) based on the statistical method. Through utilizing the Sigmoid function, the classification result of SVM will be used to fit the emission probability of binary hidden markov model (HMM) words, and realized the integration of SVM with binary HMM. The extraction precision of this model is superior to single HMM and SVM, but it did not consider the dynamic updating of extraction model.

doi:10.4304/jmm.8.6.732-738

To improve the extraction precision and of the document metadata and the adaptability of statistical learning, we presents a metadata extraction Approach of documents based on measurement-level fusion(MEAPMF) on the basis of the existing statistical learning methods.

II. Related Work

A. HMM

HMM consists of two layers: the observable layer and the hidden layer. The former is an observation sequence to be identified, and the latter is a markov process, in which each state transition carries the transition probability. The HMM model can be viewed as a quintuple as $\{S,V,A,B,\Pi\}$:

1) S represents the state set of the model, it could be denoted by $S=\{s_1,s_2,\cdots,s_n\}$;

2) V represents the set of state output symbols in the model, m represents number of symbols, the set is denoted by $V=\{v_1,v_2,\cdots,v_m\}$;

3) Π represents the probability distribution of initial state, i.e., $\Pi=\{\pi_1,\pi_2,\cdots,\pi_i,\cdots,\pi_n\}$, and π_i is the probability of s_i which is designated as initial state;

4) A represents state transition probability matrix, i.e., $A=\{a_{ij}\},1\le i\le n,1\le j\le n$, and a_{ij} is the probability of s_i transfer to s_j;

5) B represents state output transition probability matrix, i.e., $B=\{b_{ik}\},1\le i\le n,1\le k\le m$, where b_{ik} is the probability of obtaining v_k from s_i.

B. CRF

The idea of Conditional Random Fields (CRF) [12] mainly comes from the maximum entropy. CRF can be regarded as an undirected graph model or a markov random field. It is suitable to label sequenceized data in the field of natural language processing. CRF model [7] defines the word sequences $\{w_i\}$ (i=1,…,n) in the given text and the conditional probability value of the labeled sequence $\{t_i\}$ (i=1,…,n). The calculation formula (1) is presented as follows:

$$P(t\mid w)=\frac{1}{Z_0}\exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_j f_j(t_{i-1},t_i,w,i)\right) \quad (1)$$

In this formula, Z_o is a normalized factor vector of all state for the input observation sequence, and denotes the scores of all possible sequences. The calculation formula (2) is as follows:

$$Z_0=\sum_{t}\exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_j f_j(t_{i-1},t_i,w,i)\right) \quad (2)$$

where $f_i(t_{i-1},t_i,w,i)$ is feature function, it is usually assigned the boolean value. The feature feature function can be viewed as a measurement of the state transition ($t_{i-1}\to t_i$) and the whole observation sequence w. λ_j is the weight of corresponding feature function by training the model.

C. SVM

Recently, SVM has been widely applied in metadata extraction. The main idea of SVM is to transform data that cannot be classified linearly in low-dimensional space into high-dimensional space by kernel function. And the corresponding hyperplane could be found linearly classify them. Suppose the training sample setin binary classification problems is $(x_i,t_i)(i=0,1,\cdots,n)$, in which x_i is the feature vector of word w_i, and $t_i\in\{+1,-1\}$ is the corresponding class identifier. SVM classifier provides the decision function $f(x)$ composed of the input feature vector x, its purpose is to predict the t class of the unknown sample x. The optimal classification function (3) is described in [13]:

$$f(x)=sign\left(\sum_{Z_i\in SV}a_i t_i K(x,Z_i)+b\right) \quad (3)$$

If x is predicted as a positive example, $f(x)=+1$, vice versa $f(x)=-1$. In the formula, a_i is a nonzero coefficient, Z_i is a support vector, t_i is the corresponding class identifier of x, while $K(\bullet)$ is the kernel function. Undivided linear samples will be mapped to a high-dimensional space through SVM to make the samples dividable.

III. Proposed Approach

A. Framework

In order to improve the extraction precision and the adaptability of statistical learning methods, this study proposes a document metadata extraction framework based on measurement fusion on the basis of HMM, SVM and CRF.

This framework mainly consists of three functional modules such as training, testing and updating module. First, training module preprocesses the beginning part of PDF documents, selects the features of the beginning part, and then generalizes the selected features. Second, train HMM, SVM and CRF with labeled training set to generate corresponding extraction model. Testing module firstly preprocesses the beginning part of documents and generates text files, then models the posterior probability Pi (i=1,2,3) of all types of metadata by three extraction models, and realizes the decision of metadata extraction results by adopting the sum rule. In the step of updating module, by setting time quantum and number threshold value of documents, we first decides the documents that cannot be extracted properly on the basis of the metadata series extracted in the test stage, stores them in a temporary document library, and then counts the numbers of documents in the temporary document liabrary employing a statistical method based on time quantum. When the number reaches the threshold value, the model updates the training set and retrains the three statistical

learning methods so as to realize the updating of the extracted models.

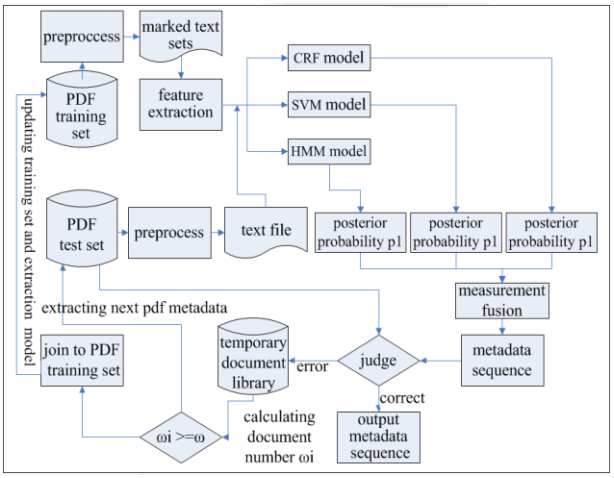

Figure 1. Metadata extraction framework of documents based on measurement fusion

B. Feature Extraction

For the purposes of feature selection, the beginning part of PDF documents should be preprocessed. First of all, using open-source tool PDFBOX to convert PDF documents page to text format; Then, using regular expression to reduce the redundant information in the text (like date of publication, journal name, journal number, details of the author, URL and Email, etc.), and to generate the target information. On this basis, the target information will be divided into 4 blocks (block1, block2, block3, block4) by adopting segmentation technology, and the end of the last block often ends with information like '1 Introduction' or the end of the first page.

1) Feature Selection

According to the problems in the process of metadata extraction of PDF documents, the following aspects should be taken into consideration in feature selection.

1) Local text features: refers to special features in spelling and forms of character as well as layout features of parts (may be a word or the metadata to be extracted) of the text sequence. For example, the abbreviated form of authors' names, the beginning and end of a line, etc.

2) External dictionary features: build a name dictionary based on 8441 first names and 19613 last names collected by Bob Baldwin and a field dictionary. For example, word lists of classes from the training data like abstract, keyword and introduction frequently appear in each line. The word frequency threshold is used to define the size of the lists, such as metadata class of abstracts contains the word "abstract" and class of keywords contains the word "keywords".

3) State transition features: feature function $f_i(t_{i-1}, t_i, w, i)$ integrates features of data sequence w and state transition $t_{i-1} \to t_i$. If state of t_{i-1} is the title, state of t_i is the author, and w_i is the names and w_i is in the first name dictionary of name database, then the value of feature function is 1. This features is only suitable for CRF model.

2) Features Generalization

Features of the training set are generalized by solving the following steps:

1) Repalece the character string in the data with the name database (replacing the string "Seymore K." with the word "name");

2) Count the numbers of high-frequency words attributing to each class in the new data and build a field dictionary.

3) Further generalize features of the data using the field dictionary similar to step (1).

Step (1) and step (3) are only needed to generalize testing features set .

C. Measurement Fusion of Three Statistical Learning Models

The measurement fusion of statistical learning model refers to combining the output results of several learning models to make decision. In order to obtain abundant fusion information, the output results of the models are formed by models of which the probability metrics are of all kinds. The more the information that measurement fusion outputs, the better performance the extraction model could achieve. Features of three statistical learning models are adopted in this testing document to model posterior probability of metadata, and finally fusion method of sum rule [14] is used to realize the final decision of metadata in documents.

1) Posterior Probability Modeling

1) Posterior probability based on HMM

When extract metadata by using HMM, the beginning parts of documents must be firstly preprocessed and blocked roughly adopting the method of text segmentation, and then the documents are subdivided in each line, each piece is represented as a state.

Mark the state of each for processed line sequences, and generate training set. ML is applied to compute the initial state distribution Π, state transition matrix A and state output matrix B to achieve the parameters of HMM model. According to the model parameters, the posterior probability of discrete symbol v_k within the i-th state s_i for the testing sample, the formula (4) is as follows:

$$P(s_i \mid v_k) = \frac{a_{ij} b_{ik}}{\sum_{i=1}^{n} a_{ij} b_{ik}} \tag{4}$$

2) Posterior probability based on SVM

When using SVM model to extract metadata, each metadata can be viewed as a class, so that extraction of metadata converts to the problem of classifying each document block. Firstly, preprocess the beginning parts of documents, extract features within each block, and establish the corresponding feature vector for each line; Then, classify the testing set. On the basis of the classification results, modify the original feature vector, and then repeatedly classify the modified feature vector with the iterative SVM classifier.

After feature selection and classifier training, text line sequences $O_1, O_2, \cdots, O_L$ are clssified into branch

sequences $U_1, U_2, \cdots, U_L$ using a one-to-many SVM classifier. Then use the Sigmoid function to convert the output distance of SVM multiple classifier to the corresponding posterior probability, and the formula (5) is as follows:

$$P(q \mid \sigma) = \frac{1}{1+\exp^{A \cdot O(\sigma)+B}} \quad (5)$$

In the formula, q denotes a certain state, $O(\sigma)$ denotes a certain word in this state. And parameters A and B are dynamically adjusted by using different training sets.

3) Posterior probability based on CRF

When using CRF model to extract metadata, the beginning parts of documents should be preprocessed, and each piece should be blocked in line. Each block represents a state; features of each block should be extracted. Then the training text will be blocked, and states of the blocked text will be marked. Using L-BFGS (Limited memory BFGS) algorithm to get parameters of CRF model from the marked training set. According to the model parameters, for the testing set, the posterior probability of text word w_k within the i-th state t_i, the formula (6) is as follows:

$$P(t_i \mid w_k) = \frac{\exp\left(\sum_{j=1}^{m} \lambda_j f_j(t_{i-1}, t_i, w_k)\right)}{\sum_{t} \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_j f_j(t_{i-1}, t_i, w, i)\right)} \quad (6)$$

In the formula, $f_i(t_{i-1}, t_i, w, i)$ is a feature function, it can be emplyed to measure the state transition $t_{i-1} \to t_i$, the whole observation sequence w and all aspects of the current procedure. λ_j is the weight of the corresponding function after training the model.

2) Derivation of Sum Rule

Suppose there are n metadata classes $\{W_1, \cdots, W_k, \cdots, W_n\}$, and R extraction models $\{M_1, \cdots, M_i, \cdots, M_R\}$ be used. Set $\vec{X}_i$ is the measurement vector that the i-th extraction model for metadata that is of W_h class, and marked as Φ. If the measurement vector of metadata of W_h class is $\vec{X}_i$ (i=1,2,…,ℝ), the posterior probability is maximum, then $\Phi = W_h$. That is:

If:

$$\begin{aligned} &P(\Phi = W_h \mid \vec{X}_1, \vec{X}_2, \cdots, \vec{X}_R) \\ &= MAX_{k=1}(P(\Phi = W_k \mid \vec{X}_1, \vec{X}_2, \cdots, \vec{X}_R) \end{aligned} \quad (7)$$

then: $\Phi = W_h$

The above rules are used to take advantage of all the valid information to obtain a decision, to compute various hypothetical possibilities by all metrics. However, it is not a feasible scheme, a posterior probability function works via a joint probability density function $P(\Phi = \vec{X}_1, \vec{X}_2, \cdots, \vec{X}_R \mid W_k)$ of high-dimensional metric statistics. In order to simplify the rules, individual extraction model output is indicated by the measurement vector $\vec{X}_i$. For posteriori probability $P(\Phi = W_h \mid \vec{X}_1, \vec{X}_2, \cdots, \vec{X}_R)$, the following formula (8) can be given by the bayesian theory:

$$P(\Phi = W_h \mid \vec{X}_1, \vec{X}_2, \cdots, \vec{X}_R) = \frac{P(\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_R \mid \Phi = W_h)}{P(\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_R)} \quad (8)$$

For unconditional joint probability density function $P(\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_R)$, the following formula (9) can be given according to the conditional distribution.

$$P(\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_R) = \sum_{i=1} P(\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_R \mid \Phi = W_i) P(W_i) \quad (9)$$

Since statistical calculations of metrics $\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_R$ are independent for each other, consequently there is the following formula (10) :

$$P(\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_R \mid \Phi = W_h) = \Pi_{i=1}^{R} P(\vec{X}_i \mid \Phi = W_k) \quad (10)$$

Put formula (8), (9), (10) into formula (7) and get the following calculation formula (11) based on posterior probability generated by each extraction model.

If:

$$\begin{aligned} &P^{-(R-1)}(W_h) \Pi_{i=1}^{R} P(\Phi = W_k \mid \vec{X}_i) \\ &= MAX_{k=1}^{n} P^{-(R-1)}(W_h) \Pi_{i=1}^{R} P(\Phi = W_k \mid \vec{X}_i) \end{aligned} \quad (11)$$

Then: $\Phi = W_h$

The above rules mixes with each individual extraction model by means of product, and generates posteriori probability to evaluate the possibility of a certain hypothesis. The rule is sensitive to a single extraction model, as long as a low posterior probability of a metadata of correct class is reported, the total output will be nearly zero. In order to solve this problem, in the process of expressing posterior probability $P(\Phi = W_k \mid \vec{X}_i)$, a parameter $\delta_{k,i} << 1$ is introduced, hence:

$$P(\Phi = W_k \mid \vec{X}_i) = P(\Phi = W_k)(1 + \delta_{k,i}) \quad (12)$$

Put formula (12) into the item on the right of formula (11) to get the following formula:

$$P^{-(R-1)}(W_k) \Pi_{i=1}^{R} P(\Phi = W_k \mid \vec{X}_i) = P(W_k) \Pi_{i=1}^{R} (1 + \delta_{k,i}) \quad (13)$$

Ignoring the quadratic terms and higher order terms in the right item of (13), then:

$$P(W_k) \Pi_{i=1}^{R} (1 + \delta_{k,i}) = P(W_k) + P(W_k) \sum_{i=1}^{R} \delta_{k,i} \quad (14)$$

Put (12) and (14) into (11) to get the sum rule applied in this document.

If:

$$\begin{aligned} &(1-R)P(W_h) + \sum_{i=1}^{R} P(W_k \mid \vec{X}_i) \\ &= MAX_{k=1}^{n}[(1-R)P(W_k) + \sum_{i=1}^{R} P(W_k \mid \vec{X}_i)] \end{aligned} \quad (15)$$

Then: $\Phi=W_h$

3) Fusion Decision based on Sum Rule

Four metadata classes to be extracted in this paper are expressed by W = {Title, Author, Abstract, Keywords}, and three extraction models are represented as M={HMM, SVM, CRF}; then the output results of sample x in training set X is the posterior probability of sample x with regards to each class. Its value can be obtained from formula (4), (5) and (6).

For testing set Y, y_m presents the metric of the *m-th* extraction model measuring the *j-th* metadata, and it is marked as Φ.

If:

$$\begin{aligned} &-2p(W_j)+\sum_{m=1}^{3}P(W_j \mid y_m) \\ &=MAX_{k=1}^{4}[-2P(W_k)+\sum_{i=1}^{3}P(W_k \mid y_m)] \end{aligned} \tag{16}$$

Then: $\Phi=W_h$

The prior probability can be obtained from the historical data, and then the final label sequence can be determined by the decision rule (16). The label sequence is namely classes of metadata, metadata extraction of documents can be realized.

D. Dynamically Updating of Three Extraction Models

As time goes on, there will be an increasingly number of metadata that cannot be correctly extracted. In order to solve this problem, a method based on time quantum is used to dynamically update three extraction models. Firstly, we set the threshold ω of the numbers of documents, use ω to represent the maximum numbers of documents that are incorrectly extracted, and the threshold is used to determine whether it is retrained or not. Then, on the basis of the threshold, dynamic updating of the model can be realized through the following steps.

1) Set the initialization time T_0 as the start of the extraction, and set T_1 as the end of the first time quantum. In the process of metadata extraction, once the extraction of a paper is done, it will be checked with an unsupervised discriminant method. The basic idea is to compare the extracted authors with the author's information dictionary, if the extracted information accords with the dictionary information, the extraction results of the authors are correct; If the extracted content before the author, then it can be regarded as the title; If the extracted content contains words exist in abstract dictionary and key words dictionary, we can regard the extracted content as abstract and key words; If the above information is correct, then the extraction is correct and the metadata can be obtained. Otherwise, it can be judged as incorrect. Then we store the documents in the temporary document library, and calculate the numbers ω1 of documents stored in the temporary document library in the time quantum $\Delta t1=T_1-T_0$;

2) If ω1<ω, then move to next time quantum calculate ω2;

3) If ω1>=ω, then three extraction models need to be updated. Mark the new document in manual way, select features according to the features of the new document, and add them to the original features library. The marked documents and the original training set, as new training set, train three statistical learning methods to generate new extraction model, go on with next metadata extraction;

4) When we arrive at T_i (i>=2), then count ∑ωi the total numbers of documents stored in temporary document library within$\Delta t1$-$\Delta t2$;; If ∑ωi>=ω, execute step (3); If ∑ωi<ω, move to next time quantum.

IV. Experimental Results and Analysis

A. Experimental Data

In order to verify the performance of the algorithm under different conditions, we conducted the following experiments. In the first experiments, the beginning parts of 935 computer research papers marked with HTML are used. The dataset is provided by CORA search engine development team in Carnegie Mellon University, and it is set to D. The datasets used in other two experiments consist of PDF papers downloaded from different OA journals sites. We totally downloaded 5300 papers from 157 OA journal sites, and deleted 1300 of them that cannot be converted (due to limited conversion tool, PDF papers in picture format, encrypted PDF papers and nonstandard PDF papers can't be converted), and the rest 4000 articles are used in the experiments. Another 400 papers which are different from others in format are also ruled out. We randomly selected 1000 papers form the rest 3600 to form dataset F, and the rest 2600 papers and 400 ruled out form dataset S. The distribution of the datasets S and F is in a cross validation way, namely 90% as the training seta, 10% as the testing set.

B. Evaluation Index

To evaluate the performance of the metadata extraction method, we adopt the precision rate P of each type of metadata as the evaluation index. Its calculation formula (17) is as follows:

$$P=\frac{Number\ of\ correctly\ extracted\ matadata}{Total\ of\ a\ certain\ type\ of\ metadata}\times 100\% \tag{17}$$

Meanwhile, in order to ensure the effectiveness of the output precision value, we adopt J fold cross-validation in the second and third experiments. Divide the dataset into J parts, use $J-1$ parts as the training set, 1 part as the testing set. Corresponding extraction precision values can be obtained from each experiment, take average value as $\overline{P}$ for J times its formula(18) is as follows:

$$\overline{P}=\frac{\sum_{j=1}^{J}p_j}{J} \tag{18}$$

In order to further reduce the error rate, J fold cross-validation is conducted for another H times, and take the average value as P_{avg}, its formula (19) is as follows:

$$P_{avg} = \frac{\sum_{h=1}^{H} \overline{P_h}}{H} \quad (19)$$

H and J both are set as 10.

C. Comparison of Precision of Metadata Extraction Methods

To evaluate the extraction precision of all kinds of metadata, we randomly selected 500 papers from the dataset D as the training set, the rest 435 as the testing set. Compare the method proposed in this paper with the following four mehods: HMM in reference [4], CRF in reference [7], SVM in reference [9] and SVM + BiHMM in reference [11]. Table.1 shows comparison of metadata extraction precision for five methods with dataset *D*.

TABLE I. COMPARISON OF METADATA EXTRACTION PRECISION FOR FIVE METHODS WITH DATASET *D*

Metadata	HMM	SVM	CRF	SVM+BiHMM	MEAPMF
Title	93.7%	94.3%	93.2%	95.4%	97.2%
Author	94.1%	92.4%	93.4%	94.5%	96.3%
Abstract	96.6%	95.8%	97.5%	99.1%	99.8%
Keywords	94.9%	93.3%	95.8%	98.9%	99.2%

In order to further verify the effectiveness of MEAPMF, dataset *F* is used to compare MEAPMF with the above four methods with the *J* fold cross-validation for *H* times. The results are shown in table.2.

TABLE II. COMPARISON OF METADATA EXTRACTION PRECISION FOR FIVE METHODS WITH DATASET *F*

Metadata	HMM	SVM	CRF	SVM+BiHMM	MEAPMF
Title	89.8%	93.4%	92.1%	94.7%	96.8%
Author	92.6%	90.3%	91.5%	93.2%	95.2%
Abstract	93.7%	92.8%	95.7%	97.6%	99.3%
Keywords	92.0%	89.6%	93.9%	98.1%	98.5%

Table. 1 and table. 2 show that the extraction precision of HMM, SVM and CRF are basically not much different from each other, the precision of the mixed extraction method SVM + BiHMM is obviously higher than that of the other three single extraction methods, while the precision of MEAPMF method proposed in this paper is the highest. It is indicated that the single extraction methods have their own advantages and at the same time have disadvantages, which cause the extraction precision to be relatively low; The SVM+BiHMM method possesses the advantage of SVM and HMM, thus its extraction precision could be improved; And MEAPMF method proposed in this paper by using the posterior probability of HMM, SVM and CRF, rules out the final results via sum rule. The result fusion of measurement can fuse the extraction information extracted by other models, so that the extraction precision is improved greatly.

It can be seen from table.1 and table.2, metadata extraction precision of various metadata in table.1 is significantly higher than that in table.2, the reason is that dataset used in table.1 consists of neat texts, in which there is no unreadable codes in the process of conversion, and thus the extraction is relatively effective. And the dataset used in table.2 consists of PDF papers downloaded from OA journal sites. In the conversion process, some unreadable codes and format disorder may arise inevitably.

In addition, according to the keywords extraction precision shown in table.1 and table.2, HMM, SVM and CRF are significantly different from MEAPMF in extracting abstract and key words. In the process of extracting, extraction of abstract, keywords, title and the authors' information is in the same way (extract in lines). Although, compared with title and authors' information, abstract and key words carry more feature information, thus the extraction precision is relatively high. But due to limitation of statistical models, the metadata extraction precision is restricted to a certain extent. For example, HMM can't use context features, limits the selection of features; CRF can't identify low-frequency feature words and misidentify high-frequency feature words; By using a few support vector to determine its decision function, SVM can avoid curse of dimensionality. But at the same time leaves out some useful samples. In addition, some abstracts and key words usually end with one or two words carrying little feature information, which are easily to be misjudged. However, by using the posterior probability of three learning methods, we fuse three learning methods through the sum rule to utilize the advantages of learning methods, thus the extraction precision is improved greatly. For example, HMM realizes extraction of text information with state transition probability and word emission probability. It avoids the phenomenon of missing low-frequency feature words and misidentifying high-frequency feature words; CRF normalizes all the features and acquires global optimization, and it does not need independence hypothesis; SVM not only considers the independent features, but takes context information into account.

D. MEAPMF Adaptability Evaluation

To verify the adaptability of MEAPMF, we adopt dataset *S* and divide it into 10 equal parts randomly. Suppose a time quantum Δt is needed to extract metadata of each part, therefore three are 10 time quantum. In this experiment, MEAPMF is firstly divided into two groups, in which one group does not involves the updating of three extraction models. That is to set ω to infinity (ω is set to 3000 in this paper), i.e., MEAPMF without updating. For another group, three extraction models need to be dynamically updated, and threshold of document number is taken 100, namely, MEAPMF with updating.

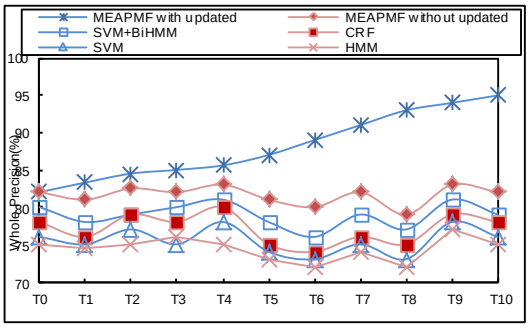

Figure 2. Comparison of the whole extraction precision for six approaches

It can be seen from Fig. 2, in the process of extracting, the experimental data are randomly added to documents

with new features. Since MEAPMF without updating, HMM, SVM, CRF, and SVM+BiHMM do not have the function of updating extraction models, the extracting precision in the whole process are relatively poor. While for MEAPMF with updating, the extraction precision increases by a lot in the process of extracting. At T_3, we count the number of papers that are not extracted correctly, then update the training set and retrain three statistical learning methods to generate new extraction models, thus extraction precision of MEAPMF with updating is improved. At T_8, MEAPMF with updating further updates the training set, and retrain three statistical learning methods, thus updates the corresponding extraction models, brings it back to a state of high performance. The updated MEAPMF can handle documents with new features, and possesses more adaptively.

V. CONCLUSIONS

Documents metadata extraction is one of the important contents of digital library construction. To deal with the eliminating defects and deficiencies existing in extraction methods, a metadata extraction method based on measurement fusion is proposed on the basis of HMM, SVM and CRF. This method model the posterior probability of various metadata based on features of HMM, SVM and CRF, and uses the sum rule to realize the final decision of extraction results. In addition, three extraction models are dynamically updated according to the time quantum and document number threshold. Compared with the existing extraction methods, the proposed method not only greatly improves the extraction precision, but has a strong adaptability and robust. In the future research, we will study more effective fusion methods and dynamic updating strategies with higher adaptability, to further improve the documents metadata extraction performance.

REFERENCE

[1] Councill I G, Giles C L, Iorio E D, Gori M, Maggini M, Pucci A. Towards Next Generation CiteSeer: *A Flexible Architecture for DigitalLibrary Deployment.* ECDL 2006 pp. 111-122.

[2] Zubair M, Flynn P, Zhou L, Maly K. Automated Template-Based Metadata Extraction Architecture. *ICADL* 2007, LNCS 4822 pp. 327-336.

[3] Seymore K, McCallum A, Rosenfeld R. Learning hidden harkov model structure for Information Extraction .*Working Notes of the AAAI Workshop on Machine Learning for Information Extraction,* AAAI Press, 1999 pp. 37-42.

[4] Liu Y Z, Lin Y P, Chen Z P. Text information extraction based on hidden harkov model. *Journal of system simulation,* 2004, 16 (3) pp. 507-510.

[5] Ojokoh B, Zhang M, Tang J. A trigram hidden markov model for metadata extraction from heterogeneous references. *Information Sciences*, 2011, 181(9) pp. 1538–1551.

[6] Zhou S X, The research of text information extraction model and algorithm. *Doctoral dissertation of Hunan university*, 2007.

[7] Peng F C, McCallum A. Accurate information extraction from research documents using conditional random fields. In: Dumais S, Marcu D, Roukos S, eds. Proc. of the Human Language Technology Conf. and North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004). New York: ACM Press, 2004 pp. 329-336.

[8] Lin S, Ng J P, Pradhan S, Shah J, Pietrobon R, Kan M Y. Extracting formulaic and free text clinical research articles metadata using conditional random fields. Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, Association for Computational Linguistics Press, Los Angeles, California, 2010 pp. 90-95.

[9] Han H, Giles C L, Manavoglu E, Manavoglu E, Zha H Y. Automatic document metadata extraction using support vector machines. International Conference on Digital Libraries, *IEEE Computer Society Press*, Washington, DC, 2003 pp. 37-48.

[10] Marinai S. Metadata Extraction from PDF Documents for Digital Library Ingest. 10th International Conference on Document Analysis and Recognition, *IEEE Computer Society Press, Washington, DC*, 2009 pp. 251-255.

[11] Zhang M, Yin P, Deng Z H, Yang D Q. SVM+BiHMM:a hybrid statistic model for metadata extraction. *Journal of software*, 2008, 19(2) pp. 358-368.

[12] Lafferty J, Pereira F, McCallum A. Conditional random fields:probabilistic models for segmenting and labeling sequence data. *Proc of 18th Int Conf on Machine Learning. San Francisco: AAAI Press*, 2001 pp. 282-289.

[13] Campbell C, Ying Y M. Learning with Support Vector Machines. Synthesis Lectures on Artificial Intelligence and Machine Learning#10, Morgan & Claypool Publishers, 2011.

[14] Kittler J, Hatef M, Duin R P W, Matas J. On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(3) pp. 226-239.

Junmin Zhao, born in Henan Pingdingshan in the middle of China, September 1978. He is a teacher of Henan University of urban construction. In 2008 graduated from Henan University of economics and law, major in computer science, received a master's degree in engineering; now study in central China normal university to get computer PhD.

Huazhong Liu, born in Henan Pingdingshan in the middle of China, July 1982. In 2012 graduated from YanShan university, major in computer science, received a master's degree in engineering; Now work in Beijing mapbar technology co. Ltd.