A Structured Information Extraction Algorithm for Scientific Papers based on Feature Rules Learning

Jianguo Chen
Fujian University of Technology /Fujian, Fuzhou, China
Software School, Hunan University /Hunan, Changsha, China
Email:cccjianguo@163.com

Hao Chen
Software School, Hunan University /Hunan, Changsha, China
Email: chenhao@hnu.edu.cn

***Abstract*—Traditional scientific papers are unstructured documents, which are difficult to meet the requirement of structured retrieval, statistical classification and association analysis and other high-level application. Hence, how to extract and analyze the structured information of the papers becomes a challenging problem. A structured information extraction algorithm is proposed for unstructured and/or semi-structured machine-readable documents. With extracted rules after feature learning on the basis of analyzing the basic structure and format features of traditional scientific papers, the proposed scheme extracts the title, author, abstract, keywords, text and other elements of paper from the unstructured documents. Then the proposed scheme exports the structured text from the traditional scientific papers with the format required by multi-dimensional scientific papers, which can meet the requirements of structured retrieval, statistical classification and other high-level applications of scientific papers.**

***Index Terms*—Information Extraction, Feature Rules, Multi-dimensional scientific Papers**

I. INTRODUCTION

Scientific papers are important output forms of scientific and technological research activities and an important means to reflect scientific researches achievements and carry out academic exchange [1]. With the rapid development of modern multimedia technology and network communication technology, the expression forms of information technology have been developed from a single two-dimensional form to an integrated form such as audio, video, 3D animation and other multimedia expression [2]. Researchers can display dynamic, complex process of science and technology experiments with the multimedia forms in a multi-angle and vivid way, present and disseminate their academic ideas and insights all-roundly. Currently, Most of the scientific papers are edited with word processing software, such as Microsoft Word, WPS, etc. other than some efficient multi-dimensional scientific papers editing tools which support multiple media formats [3]. Although Word and other word processors support the inserting of multimedia objects in way of inserting objects, it is inconvenient in the metadata extraction and format conversion. Because Word and PDF formats are unstructured documents which cannot be directly extracted the title, author, abstract, keywords, text and other elements of the papers, it is difficult to meet the requirements of structured retrieval, statistical classification and association analysis and other high-level applications of scientific papers. We need to extract all the paper elements stored in the traditional scientific papers in form of Word document, and then convert them into multi-dimensional scientific papers stored in the structured form [4]. With this method, the application of traditional scientific papers in the future scientific research can be improved, existing research achievements can be protected and the spreading of science idea academic thinking and the converting efficiency scientific research can be speeded.

According to different implementation method, the Information extraction technology can be divided into a dictionary-based text information extraction [5-7], text information extraction based on the Markov model and hidden Markov models [8], feature-based rules and semantic-based rules for text information extraction [9], of which the dictionary-based and Markov model-based information extraction are used mainly in Web field; only the method of feature-based rules and semantic-based rules can be applied not only to Web field but also Word, PDF documents field for text information processing.

This paper develops with the focus on the structure extraction of traditional scientific papers, designs and implements a structured traditional scientific paper extraction system with generate extraction rules after text feature learning, through analyzing the basic components and format features of traditional scientific paper. The system could export structured text of traditional paper from Word, PDF and other documents with the requirement of multi-dimensional technology format, and edit and read papers in multi-dimensional authoring tool.

doi:10.4304/jsw.8.1.55-62

II TECHNOLOGIES INTRODUCTION

A. Word Technology

Microsoft Word is one of the important components in the Microsoft Office integration office software, and is one of the most popular texts editing software currently. Word documents are widely used in various industries, it is most frequently used application software in the scientific literature editing for researchers.

The Word System's objects model is large and complex structure, its basic model is the basic properties of the document, text attributes, paragraph attributes, tables and pictures and other object models. The highest level object in all objects of Word is Application, followed by Documents, Document objects, the objects contained by Documents and Document object is as follow: Range, Paragraphs (Paragraph), Tables (Table) [8].

The process controls for word include the following three aspects:

(1) Basic operations: Get the selected range of text contents in Word document, and then read or set the text information through the range objects. Get the paragraph object using the index of Paragraphs sets, and then get the range of paragraphs in documents through the range property of passage object, and finally read or set the text in paragraph through the range object's Text property.

(2)Table operations: obtained the Tables collection of information through the Document object's Tables property, we can use the Table Object's Cell ( ) method to get or set each cell's text content in table.

(3) Image operations: obtained the Inline Shape collection with setting the Inline Shapes properties of Document object. Inline Shape objects include pictures, ActiveX controls and other information. We can insert picture information into a Word document using the Add Picture method of Inline Shapes objects.

B. PDF Technology

PDF is one of the most widely used document formation. It has an important advantage of cross-platform, portable, faithfully maintained the original appearance of the document display, but it can only be read by special programs.

The basic structure of PDF document consists of four parts:

Header: mark the PDF specification version which PDF document followed;

Body: The object contained in the document which composition of the document;

Trailer: the file location cross-reference table for and the position of some special object in the body.

Fonts in PDF document are in the form of program. Font program using a dedicated language to write, content stream on the page describe the font by specifying the font dictionary and string object, the string object is interpreted as a identifying the font in the font of one or more character code sequence .

C. XML Technology

XML is a rule language that defines the semantic markup, which provides a format for describing structured data, and achieve the separation of the data storage and software interface. Because of its self-describing, platform-independent, semi-structured, etc., XML has been widely used in scientific data conversion and information retrieval fields, is a new standard for data exchange. XML features are as follows:

(1) Well-formed: XML document is a structured format file and is composed by markup, and is one of the best versatility data format in the computer field.

(2) Platform-independence: XML document independents of platforms and applications, achieve the data transfer and exchange between different users and processes, and unrelated with the operating system type.

(3) Self-descriptive: XML is a semantic markup language that only contains the mark which describes the content of the document but not the appearance format of the document.

(4) Scalability: XML is a meta-markup language that only defines a set of rules for semantic markup.

III. INTELLIGENT RECOGNITION AND EXTRACTION ALGORITHM OF ELEMENTS IN TRADITIONAL SCIENTIFIC PAPERS

As for the information extraction of Word documents, XML supporting technology of Word 2003 is used to transferring objects; and it carries out the machine learning and intelligent extraction of each paper elements in the Word document and then generates XML document [10, 11]. Word document adopts a self-description definition method, which can describe the font, word size, paragraph number and other displayed information of the object before defining the object [12]. Word 2003 provides XML interference itself and it can generate a corresponding relationship between the paper elements and XML format criterion; meanwhile, the display feature information of each text, picture and table in the Word document is the source and basis of information extraction rules.

As the content information and displayed information of PDF document are stored in different objects, and the object content has many programming method in PDF document [13]. It must be converted into XML document and carries out the information extraction and the sample learning modules with SAX technology.

With the above analysis, we determine that the orientate position based on feature is the key technology of information extraction from Word, PDF document [14]. The text information of Word, PDF document has many features, which includes format feature: font, word size and over striking of paper elements; content feature: whether it is fixed text, whether it contains feature text, start mark and end mark. These features will be used as displayed features of text information in the creation of extraction rules.

A. General Idea of the Algorithm

With the combination of information extraction, XML structured technology and machine learning technology, this algorithm realizes the conversion of traditional scientific papers stored in the form of Word document into structured multidimensional scientific papers based on XML metadata criterion. The general operation flow chart of the algorithm is shown in Figure 1.

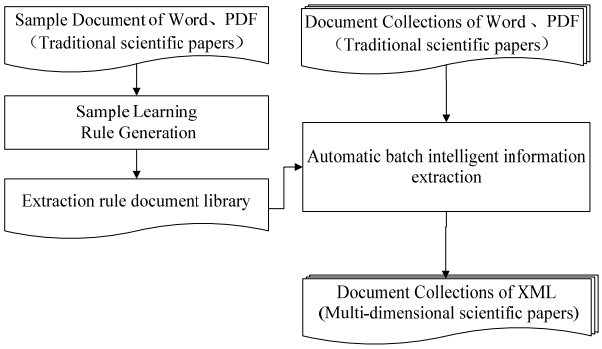

Figure 1. The flow chart of information extraction based on feature rules.

The general handling process of the algorithm is as the follows:

Firstly, sample learning: because the paper publishing formats of each journal are not all the same, it must carry out the sample learning of traditional papers in different journals [15]. From the research result of the last chapter, it is known that the structure of the scientific papers is made up of title, author, the working department, place, and e-mail of the author, abstract, key words, content, main parts, references, acknowledgement, appendix and other basic elements [16]. This algorithm will mark the text, format feature of each paper element in the Word, PDF document, combine the paper structure, generate the extraction rules and store them in the rule base.

Secondly, information extraction: the user chooses the paper document to be converted from the traditional paper document base, at the same time, chooses the corresponding extraction rule document from rule document base. The system will read the content of the original Word PDF document through document object, transfer information extraction algorithm, and carry out the intelligent recognition and feature extraction of the paper elements with a rule parser, and then get each key element of the paper.

Lastly, creating multidimensional paper: the system read the metadata rule document of the multiple dimension paper through document object, transfer XML parser and generate XML based structured multi-dimension scientific papers which in accordance with multiple dimension paper storing format with the obtained key elements of the paper.

B. Sample Learning and Extraction Rules Generation Algorithm

The ample learning and extraction rules generation algorithm is shown in Figure 2.

1) Sample learning

Because the paper requirements in each journal are not all the same, the paper in each journal shall be carried out sample learning in order to extract the paper structure [17]. Firstly, a paper document with representative display shall be chosen so as to contain a maximal range. The traditional scientific papers Sample document is shown in Figure 3.

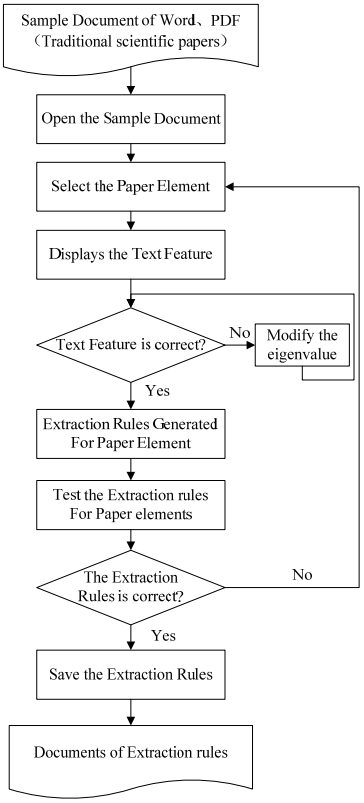

Figure 2. The flow chart of sample learning and extraction rule generation algorithm.

The user opens the paper sample document with some related editing tools, gets an overall understanding of the content and logic structure of the sample document and then analysis the paper structure. Since the sample document is representative, so is the feature of the paper elements.

Study on Method of Web Content Mining for Non-XML Documents

Jianguo Chen[1,2], Hao Chen[1,3], and Jie Guo[1]
[1]Software School, Hunan University, Changsha, 410082,China
[2]Software College,Fujian University of Technology,Fujian,350003,China
[3]School of Information Science and Engineering,Central South University,Changsha
cccjianguo@163.com, chenhao@hnu.cn

Abstract

Web content mining is an important way of Internet information collection and analysis, but most of web pages are non-XML documents, how to extract useful information efficiently from massive web pages is a interesting research topic. On the basis of analyzing the features of web content mining, a XML-based web content mining method is proposed. Firstly, it defines the authority web page using the HITS algorithms, then transforms the non-XML documents into structured XML documents after the data cleaning and extracting by HTML Tidy, finally does data mining on the XML document using text clustering techniques. A science paper

the relational database data mining[4]; study include the use of web structure mining and web mining, has not addressed the content mining; in the practical application of the elaboration is relatively simple. In this paper, research is based on XML, web content mining model, based on the key technologies, combined with the Internet, the traditional scientific papers examples of automated extraction system is proposed which is based on XML technology for the web content data, using text clustering classification techniques for surface XML document data to the data miningmethod.

2 web Mining And XML Technology

2.1 web Mining

Figure 3. Traditional scientific papers Sample document.

The user chooses the title, author, author's department, e-mail, abstract, key words and other paper elements from Word, PDF sample document. After choosing one item, the system will display the text features of paper elements, such as font, word size, over striking, text

content and other information and then studies the sample. The purpose of sample learning is to get the feature value of the each paper element data.

Algorithm formula:

$$K = \{k1, k2, k3\ldots kn\} \quad (1)$$

K is the feature collection of each paper element extraction; k_n is the text feature item of paper element, such as font, word size and over striking etc.

$$V = \{v1, v2, v3\ \ldots vn\} \quad (2)$$

V is the value collection of each paper element feature; v_n is feature value, such as Song style, small size five and over striking etc.

$$R= \{ (k_1, v_1), (k_2, v_2)\ldots (k_n, v_n)\} \quad (3)$$

R is the collection of extraction rules, for example, the extraction rule collection of paper title is R paper title= { (font, black), (size, three), (flush, middle)…}.

2) Extraction rule generation

Through the study of sample document of scientific document, we can conclude that the feature of each element of scientific papers can be divided into the following two aspects:

Format feature: the font, word size, heavy type of paper elements;

Content feature: whether it is fixed text, whether it contains feature text, start mark and end mark.

The feature information of papers is an important source and basis of extraction rules, therefore, whether the acquired feature information is correct, overall and representative directly relates to the accuracy of extraction rule and influences the extraction efficiency of information.

Therefore, the extraction rule of this algorithm is made up with [whether it is fixed text]; [whether it contains feature text]; [left mark, right mark]; [font]; [word size]; [over striking]; and these items are not always used at the same time, while most of the time, some of them are usually used together. The following part will explain the feature and corresponding extraction rules of each paper element in details. Each paper element need to set its extraction rule collection R={ (font, black), (size, three), (flush, middle)…}, then extract the information. The key flow chart of information extraction algorithm is as follows:

```
/**************************************
ALOGORITHM: Extraction Rule Generation (
Wordstr : string of word document; /*string for extraction*/
startstr: start string; /*the end target for extraction */
endstr: end string /*the end target for extraction */)
{
str ← startstr + "cccjianguo" + endstr;
RegListArea ← str.Replace ("cccjianguo", @"([\S\s]*. )");
tmpreg: Regex ← new Regex (RegListArea: string[], RegexOptions.Compiled);
sMC: MatchCollection ← tmpreg.Matches (Wordstr);
IF sMC.Count != 0 do
{
 RsltAry ← new string[sMC.Count];
 FOR I = 0 to sMC.Count DO{
  RsltAry[i] ← value of sMC[i].Groups[1];
 }
}
ELSE
  RsltAry ← NULL;
RETURN (RsltAry)
}
**************************************/
```

The information intelligent extraction part will receive the traditional scientific paper document and corresponding feature rule document of a certain journal, then transfer the information extraction function, and finally output the XML structured multidimensional scientific paper elements.

This algorithm is fit for all the documents which are not audio, video and animated, which is rather general and can classify and manage a number of documents according to the name of journal, and then carry out the feature- based information extraction.

IV. System Realization

A. System Function Modules

The system includes the following four function modules: module for the management of traditional scientific papers, module for the management and generation of sample learning and extraction rule, module for the management of extraction and module for the management of multidimensional papers, of which module for the management of traditional scientific papers includes the management of journal, traditional scientific papers, the structure of traditional scientific papers and the open of traditional scientific papers; module for the management and generation of sample learning and extraction rule includes rule generation and the management of rule document base; module for the management of extraction mainly includes automatic extraction in bulk and manual extraction; and the module for the management of multidimensional papers includes the management of multidimensional papers base and multidimensional papers' structure. system functional modules shown in Figure 4.

B. Paper Elements Learning and Extraction Rule Generation

1) Header information of the papers

Paper title: be marked with font, word size and over striking, for example, the display feature of a paper in a journal is: black, small size two and over striking.

Author: the display feature is font, word size and not in over striking; the text feature is that "author" is in the next paragraph below "paper title" and "paper title" is the begin mark of "author". As one paper may belong to many authors, there shall be a separation mark ", " to further extract the information item "author".

Author's department: "author" shall be the begin mark of "author's department", and then the end mark will be

the “end mark paragraph” of the paragraph where the author’s department is.

Email: the email character string contains a mark “@” which is related to the author, and there shall be a separation mark “; ” to further extract the information item “email”.

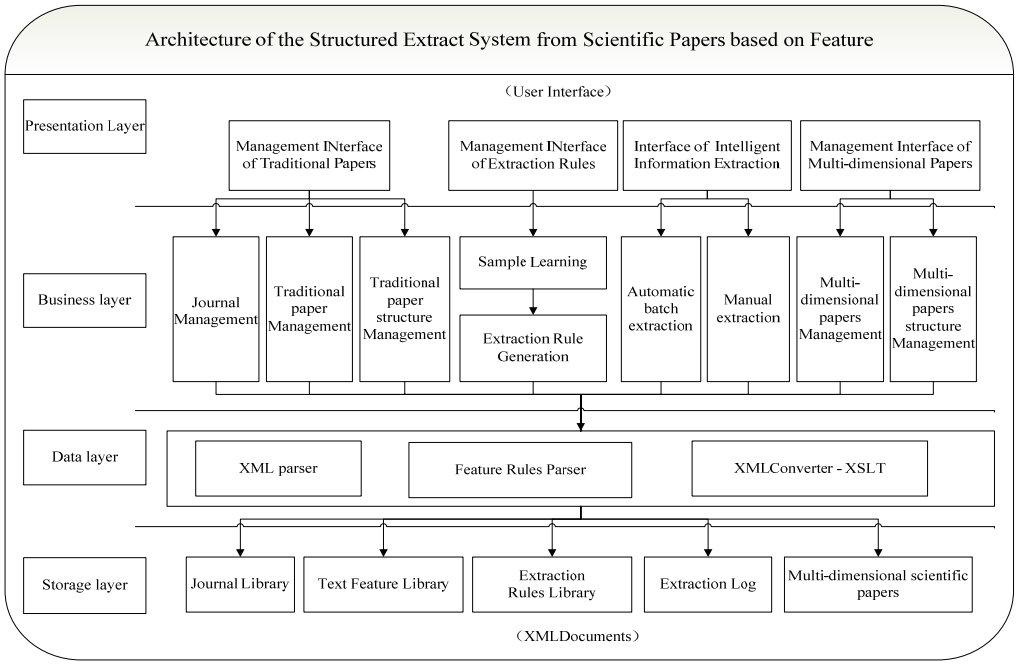

Figure 4. System Functional Modules.

Abstract: the display feature of “abstract” content is the same as that of the main content. In order to keep the accuracy of learning, we use the text feature as the extraction basis, “abstract” as the begin mark and “key word” as the end mark.

Key words: “key words” is the begin mark and “paragraph end” is the end mark. Since there are many key words, the separation mark “; ” shall be used to further extract the obtained information item “key words”.

Figure 5. Header Information Rule Generation (part).

The information extraction for header information of paper needs to set each element’s characteristic rule for the paper’s sample document manually and then to test each other. The header information of paper includes title, author, author affiliations, address, Email, Abstract, keywords, etc. The detailed function interface for header information extraction rule generation is shown in Figure 5.

2) Body information of the papers

Chapter (I): Extraction rules for Chapter (I) will be combined with the display features and text features, which can be identified by the font family, font size and font weight of the text. For example: The display features for Chapter (I) of a journal paper as follow: font family is Times New Roman, font size is 16pt and font weight is bold. The text features for then as follow: the first character is chapter number, which be formatted in {1, 2, 3, ...}, {Chapter 1, Chapter 2} or {I, II, III, …}{Chapter I, Chapter II}. So we can set start target for Chapter (I) with the "chapter number" which is the first character in the paragraph, and set end target with paragraph’s end mark.

Chapter (II) (III): the sub chapters located after chapter (I), and can be identified by the font family, font size and font weight of the text. For example: The display features for Chapter (I) of a journal paper as follow: font family is Times New Roman, font size is 14pt and font weight is bold. The text features for then as follow: the first character is chapter number, which be formatted in {1.1, 1.2, ...}, {Chapter 1.1, Chapter 1.2}. So we can set start target for Chapter (II) with the "chapter number" which is the first character in the paragraph, and set end target with paragraph’s end mark. The method of extraction for Chapter (III) is the same to Chapter (II).

Contents: extraction rules for the papers contents text will be combined with the display features and text features. For example: The display features for Chapter (I) of a journal paper as follow: font family is Times New Roman, font size is 12pt and font weight is normal.

Pictures and Tables: extraction for all the pictures and tables in the document will through the Word document object, pictures, tables, so it is easy to extract the picture ID, picture title, table ID and table title from papers;

The detailed function interface for body information extraction rule generation is shown in Figure 6; we can click and drag the mouse in the word document, and select the information block. Then click the "rule test" button which behind the rule. System extract the each element for the current paper document with extraction rule by calling the word object program, to ensure the correctness of the extraction rule.

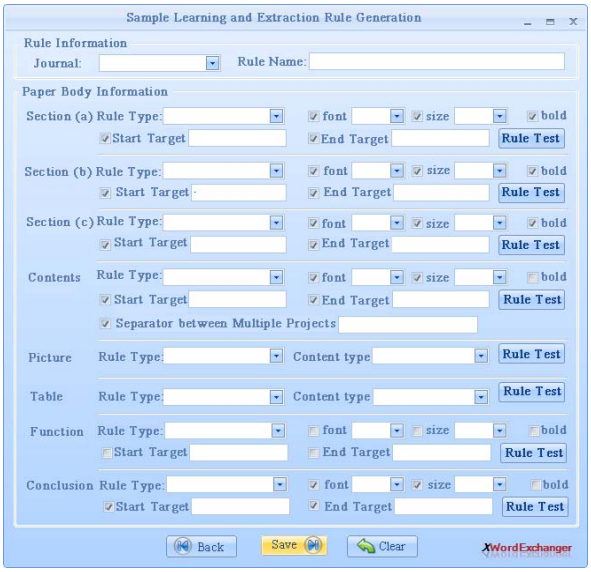

Figure 6. Body Information Rule Generation (part).

3) Ending information of the paper

References: according to the classification, the formats of references are as follows:

Journal: [number] author (family name is in front of the surname, the surname can be abbreviated, thereafter the same). Title (the first letter of the English title shall be in capital, the rest shall be lowercase): subtitle (if any). Journal name (complete name), year, volume (number): page.

Degree paper: [number] author. Paper title [D]. Department address: department name, year.

The extraction of the ending information of the paper includes: the extraction of references and appendix. Since the references can be classified into works, journal papers, degree papers, conference papers, network references and newspaper references, we shall set different extraction rule respectively for each kind of reference, and the detailed function interface is shown in Figure 7.

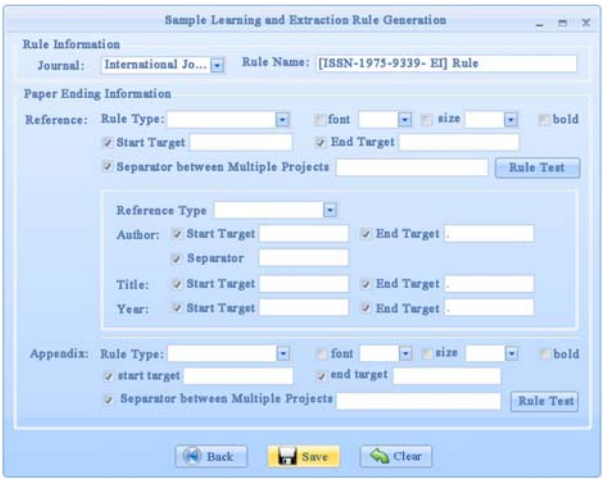

Figure 7. Ending Information Rule Generation.

After analyzing the paper structure and learning the paper element, the extraction rule of the paper elements is as follows:

The collection of extraction feature of the paper element K = {font, word size, over striking, line space, flush}

The value of the collection of each element feature value of papers V and the value of extraction rule collection R={ (k1: v1), (k2: v2), (k3: v3), …} are shown in TABLE I.

TABLE I.
TABLE OF EXTRACTION RULES GENERATED FOR PAPER ELEMENT

Paper elements	Exaction Rules
title	R-title= {(font: Times New Roman),(font-size:12), (font-weight: bold) ,(line-height: 2)(text-align: center)…}
author	R-author= {(font: Times New Roman),(font-size:8.5),(font-weight: normal)(line-height: 1.5)(text-align: center)…}
abstract	R-abstract= {(font: Times New Roman),(font-size:8.5),(font-weight: normal)(line-height: 1.5)(text-align: left)…}
Key words	R-key words= {(font: Times New Roman),(font-size:8.5),(font-weight: bold)(line-height: 1.5)(text-align: left)…}
contents	R-contents= {(font: Times New Roman),(font-size:8.5),(font-weight: normal)(line-height: 1.5)(text-align: left)…}
…	…

C. Automatic Extraction Module of Paper Elements

The task of this module is to receive the journal name and corresponding feature rule document to be extracted in bulk. The system will find the store catalogue corresponding to the journal and then with tree traversal algorithm to travel and operate the catalogue. Once the system checks the traditional scientific papers in Word, PDF form to be converted into structured multidimensional papers, it will transfer the feature rule extraction algorithm and finally output the multidimensional scientific papers in XML structured format which are in accordance with the store standards of multidimensional scientific papers.

```
/*****************************
   ALOGORITHM: ExtractionElement (
   V0: doc; /* Word, FPD Document */
   V1: searchstr /* Element text feature */)
   {
   strKey ← searchstr;
   i ← 0;
   iCount ← 0;
   Microsoft.Office.Interop.Word.Find wfnd;
   IF doc.Paragraphs != null && doc.Paragraphs.Count > 0 DO
      iCount = doc.Paragraphs.Count;
   FOR I=1 to iCount DO{ //for all paraguaphs of the document
         wfnd ← doc.Paragraphs[i].Range.Find;
         wfnd.ClearFormatting ();
         wfnd.Text ← strKey;
         IF wfnd.Execute DO
           RETURN (searchstr)
      }
   }
   *****************************/
```

The detailed function interface for Automatic batch extraction module of paper elements is shown in Figure8.

V. System Test

The author has chosen the scientific paper Study on Method of Web Content Mining for Non-XML Documents. International Conference published in the period of graduate study as the sample object to test the system. The author also automatically extracts the information of the 100 papers in Word, PDF published in the same journal.

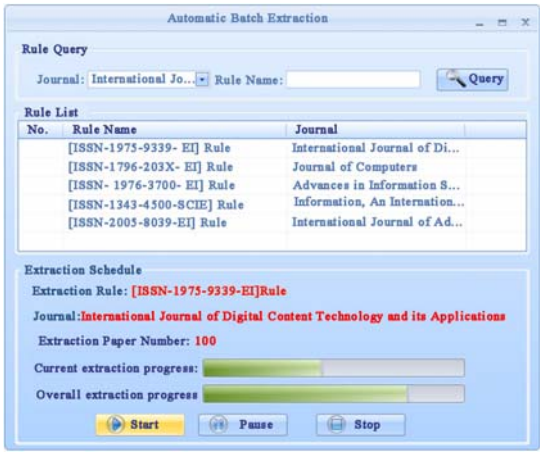

Figure 8. Automatic batch extraction.

The system circumstance is as follows: Microsoft Windows Server 2003 + .Net Framework 3.5

The test index used is the rate of overall extraction and the rate of accurate extraction.

$$P_{CE} = \frac{N_e * \mathrm{R_e}}{N_{ae} * (K_{ae} + V_{ae})^2} \tag{4}$$

P_{CE}: Probability of Complete Extraction; N_e: The Number of Elements Extracted from Papers; N_{ec}: The Total Number of Elements for Papers.

$$P_{AE} = \frac{N_{ec} * (K_{ec} + V_{ec})^2}{N_e * \mathrm{R_e}^2} \tag{5}$$

P_{AE}: Probability of Accurate Extraction; N_{ec}: The Total Number of Elements Extracted from Papers; N_e: The Number of Elements Extracted Correctly from Papers.

Test group one is the test data of the system implemented in this paper; test group two is the test data obtained from the original system in the paper Research of HTML and PDF Information Extraction Technology Based on XML; and test group three is the test data obtained from the original system in the paper Research of Semantic Information Extraction in PDF Document. The test and comparison result are shown in TABLE II.

TABLE II.
COMPARE OF SYSTEM TESTING RESULTS

Standards for Test	Group One		Group Two[6]		Group Three[13]	
Elements of Test	P_{CE}	P_{AE}	P_{CE}	P_{AE}	P_{CE}	P_{AE}
Title	98%	98%	93%	89%	96%	91%
Author	93%	99%	91%	87%	91%	95%
Author's Company	94%	98%	--	--	--	--
Abstract	95%	97%	94%	96%	93%	95%
Key Words	89%	94%	--	--	--	--
Contents	84%	81%	--	--	--	--
Pictures	98%	92%	--	--	--	--
Tables	85%	89%	--	--	--	--
References	95%	97%	--	--	--	--

From the test description, in test group two and three, only the paper title, author, email and abstract are tested, while other paper elements are not tested. The test result shows that the rate of overall extraction and the rate of accurate extraction are relatively high in the paper extraction test implemented in this paper. Of which, the rate of overall extraction and the rate of accurate extraction for the XML element title, picture, acknowledgement has reached over 95%, but the extraction result of author and email is not satisfying.

After checking the test document, it is found that what influences the rate of overall extraction is the numbering of author names in the references of 6 Word, PDF document is not correct. That is there is no separation mark ", " but "; " between many authors, which is not in accordance with the extraction rule learn from the sample documents and will cause low rate of overall extraction.

VI. Conclusion

This paper analyses and studies the information extraction technology based on feature in details as well as the format feature and basic structure of traditional scientific papers in key journal at home; in the paper, the storing standard of multidimensional scientific papers is analyzed. With the expounding of some basic theories and knowledge related to Word and PDF information extraction technology, the author proposed an information extraction algorithm of traditional scientific paper based on feature and on the basis of the above

theory designs and implemented a structured extraction system of traditional scientific paper based on feature. The test result shows that the automatic extraction will make the extraction better. The design idea of the system frame and key module as well as its combined application in scientific papers has a high reference value for the information extraction in other areas.

ACKNOWLEDGMENT

This work is partially supported by the National Science Foundation of China under Grant No. 61070194; the Information Security Industrialization Fund from NDRC of China in 2009 (No. [2009]1886); the Major Achievements Transfer Projects of MOF and MIIT of China in 2010 (No. CJ [2010] 341): Science and Technology Key Projects of Hunan Province (2011FJ2003).

REFERENCES

[1] Yanhong Zhai, *Structured data extraction from the web*. Urbana, Chicago: The University of Illinois at Chicago, 2006.

[2] Mehmet Kaya, Reda Aihajj, "Online mining of fuzzy multidimensional weighted association rules", *Applied Intelligence*, vol.29, pp.13-34, 2008.

[3] Tzu-Fu Chiu, Chao-Fu Hong, "Using text mining and chance discovery for exploring technological directions via patent data", *in 2010 IEEE International Conference on Systems Man and Cybernetics (SMC). Barcelona, Spain,* pp.3853-3860, July 2010.

[4] Srinivas Vadrevu, *Automated information extraction from web pages using presentation and domain regularities*, Phoenix, AZ: Arizona state University, 2008.

[5] Houssam Nassif, Ryan Woods, "Information Extraction for Clinical Data Mining: A Mammography Case Study", *in 2009 IEEE International Conference on Data Mining Workshops (ICDMW). FL, USA*, pp.37-42, December 2009.

[6] Bin Zhou, Yan Jia, "A Distributed Text Mining System for Online Web Textual Data Analysis", *in Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). Huangshan, China*, pp.1-4, October 2010.

[7] Sushain Pandit, *Ontology-guided extraction of structured information from unstructured text: Identifying and capturing complex relationships*, Ames, Iowa: Iowa State University, 2010.

[8] D. Carrell, D. Miglioretti, "Coding free text radiology reports using the cancer text information extraction system (caTIES)", *In American Medical Informatics Association Annual Symposium Proceedings (AMIA). Rochester, USA,* pp.889-893, September 2007.

[9] L. Rokach, O. Maimon, "Information retrieval system for medical narrative reports", *In Proc. of the 6th International Conference on Flexible Query Answering Systems (FQAS). Lyon, France,* pp.217–228, June 2004.

[10] Bojan Babic, Nenad Nesic, "A review of automated feature recognition with rule-based pattern recognition", *Computers in Industry*, vol.59, pp.321-337, 2008.

[11] Zhaohui Huang, Huajun Chen, "Semantic Text Mining with Linked Data", *In NCM '09. Fifth International Joint Conference. Seoul, korea,* pp.338-343, Aug 2009.

[12] Jose A, Rodriguez-Serrano, "Unsupervised writer adaptation of whole-word HMMs with application to word-spotting", *Pattern Recognition Letters*, vol.31, pp.742-749, 2010.

[13] Christopher C. Yang, Chih-Ping Wei, "Cross-lingual thesaurus for multilingual knowledge management", *Decision Support Systems*, vol.45, pp.596-605, 2008.

[14] Atanasova T., Kasheva M., "Analysis of the possible application of Data Mining, Text Mining and Web Mining in business intelligent systems", *In 2010 Proceedings of the 33rd International Convention. opatija, Croatia,* pp.1294-1297, May 2010.

[15] Un Yong Nahm, B.S., M.S, *Text Mining with Information Extraction*, Austin, Texas: The University of Texas at Austin, 2004.

[16] Sanchez D., Martin-Bautista M.J., "Text Knowledge Mining: An Alternative to Text Data Mining", *In IEEE International Conference on Data Mining Workshops (ICDM). Pisa, Italy,* pp.664-672, December 2008.

[17] Friedlin J., Mahoui M., "Knowledge Discovery and Data Mining of Free Text Radiology Reports", *In Healthcare Informatics, Imaging and Systems Biology (HISB). California, USA,* pp.89-96, July 2011.

Jianguo Chen born in 1985, Lecturer of Xiamen University of Technology, His main research interests include software engineering, data mining and project management.

Hao Chen born in 1975, associate professor of Hunan University, His main research interests include mobile search, Web mining and service science.