Research on PDF documents information extraction system Based –on XML

Wende Zhang Yanjuan Song

(Library of Fuzhou university ,Fujian 350002,P.R.of China)

Abstract With the development of Internet , Web has become the biggest resource warehouse. Facing the more and more abundant digital information resources, the traditional way for information management cannot catch up with the development of modern society . So a kind of systematic technology is needed urgently to manage the digital information resources .To meet with this kind of need the digital library came into being . The digital library is a new developing and promising application , among which Information Integration is a basic component. The digital library's information systems adopt the specific resource forms, For example ,the resource form of Wangfan is pdf and that of CNKI is caj . So it is unconveniencing that users have to install specific reader before reading these resources .The work we are doing is to develop a new digital library ,from which the users can read the resources directly and it is not necessary for the users to read them by the specific reader. From the point of literature management , it is necessary for us to decompose a resource both in structure and semantics . The latter is the advantage of XML. XML is a technology dependent on content and Internet flat – independent .On the basis ,we propose a metadata form which is imbedded among resources based on XML , as a result of which a new data resource format can be integrated . First. Change each kind of possible form (PDF , CAJ , RTF ,etc.) into XML , Second. Make and change XML mark into a conversion rule of HTML mark .After that ,the old format of the resources are transmitted to the new format ,which can be showed by the browser directly.

Keywords: Information Extraction ， PDF ， XML

1 Introduction

Adobe PDF-Format (Portable Document Format) is put forward by American Adobe corporation . For its excellent trait , PDF-Format has become the ideal document format in the formatted –information transmission of Internet. Recently, people become more and more likely to submit articles of science and technology in PDF -Format, such as Wangfan . However, PDF-Format is good at describing the printing format of articles, while bad at the contents of articles. As a result , it is inconveniencing for people to retrieve information . So , research on extracting information from PDF-formatted document is of great significance .

Extensible Makeup Language (XML)is the data-exchange standard proposed by W3C . XML is not only a technology dependent on content and Internet flat-independent , but also the ideal tool to handle distributed data this age . XML is content-dependent , which is just the disadvantage of PDF-Format .

The article is structured as follows . Firstly , we commit ourselves to design a self-contained Document Type Definition (DTD) , which provides a semantic frame for scientific and technological articles . After that , we illuminate the PDF -Format in brief . On the basis , we dwell on how to design a PDF information extraction system , which is able to transfer a PDF-formatted article of scientific and technology to a valid XML document according to the above DTD document .

2 Work Flow of Designing PDF Information Extraction System

2.1 DTD Designing

The first step we are doing is to contrive a DTD document that provides a semantic frame for PDF documents.

After referring to the *Simplified DocBook* which is the subset of *DocBook* element and considering the scientific and technological articles have the characteristic of normative –diction , we analyze and select two basic information to picture them .

(1) Exterior Information metadata (*ArticleInfo*) : It is the metadata depicting the exterior trait of a scientific and technological article . It contains *author*, *address* , *edition* , *bibliography* and so on . The exterior information metadatas are the vital clues for information retrieval.

```
<! ELEMENT Articleinfo ( authorgroup ， edition， bibiography) >
<! ELEMENT authorgroup (address ， author+)>
<! ELEMENT address (department， city， zip ， email )>
<! ELEMENT author (name ， birth ， sex ， degree， research)>
<! ELEMENT edition (ediname， pagenums， volumenum， issuenum， pubdate)>
<! ELEMENT bibliography(bibliodiv+) >
<! ELEMENT bibliodiv( title， biblioentry) >
<! ELEMENT biblioentry ( (authorgroup ， title ， publisher ， date) | ulink) >
<! ELEMENT authorgroup (author_name+) >
<! ELEMENT publisher (publishername， address) >
<! ELEMENT department (#PCDATA) >
<! ELEMENT city (#PCDATA) >
      … …
<! ELEMENT ulink (#PCDATA) >
<!ATTLIST ulink url CDATA>
```

(2) Interior Information metadata : It is the metadata portraying the semantic information of articles , such as *Title, Abstract* ，*Keywordset*，*Section*，*Para* . It is of great sense to retrieve articles by the interior information metadata for the reason that the efficiency of information retrieval can be improved greatly .

① *Title* : It reflects the core content of article most directly .

```
<! ELEMENT  Title (#PCDATA)>
```

② *Abstract* : It is the abstract of the aimed article .

```
<! ELEMENT  Abstract (#PCDATA)>
```

③ *Keywordset* : It is the set of keywords showed by the aimed article .

```
<! ELEMENT Keywordset ( keyword+) >
<! ELEMENT keyword ( #PCDATA) >
```

④ *Section* : It describes the sections of the aimed article . In order to improve the efficiency of article classification and retrieval , it is necessary for us to analyze the sections and chapters structure of articles. An article is composed of sections , and paragraphs constitute sections . One of the important task we have to do is to judge the themes of paragraphs and sections . The details on how to judge the themes of paragraphs and sections in an article will be amplified in 2.3.4 .

```
<! ELEMENT Section ( sect_theme，  (Section|para+) *)>
<! ELEMENT sect_theme ( #PCDATA) >
<! ELEMENT para (para_theme *) >
<! ELEMENT para_theme (#PCDATA) >
<!ATTLIST para id ID   # REQUIRED>
```

2.2 Document Format of PDF

In order to accomplish the semantic information extraction from PDF documents , it is urgent for us to understand the document format of PDF

2.2.1 PDF Objects

The basic elements constituting PDF documents are PDF Objects . PDF support seven basic object types : *Boolean* , *String* , *Name* , *Dictionary* , *Number* , *Array* , *Null* , *Stream* , among which *Dictionary* Objects are the main components of PDF documents . Page-layout and word -warehouse in PDF documents are all represented by *Dictionary* Object .

PDF Objects can be divided into *Direct Object* and *Indirect Object* . PDF *Indirect Object* is a signed object . It consists of object sign , *Direct Object* and keyword *endobj* . A great number of *Indirect Object* and *Indirect Reference* are used in PDF documents .

2.2.2 Physical Structure of PDF

Four parts make up the physical structure of PDF-Format .

Head
Body
Cross table
Trailer

(1) Trailer : File trailer contains the information such as address of cross table , address of root object *Catalog* and encryption .

(2) Cross Table : Cross table is the address reference table designed to access *Indirect Object* at random .

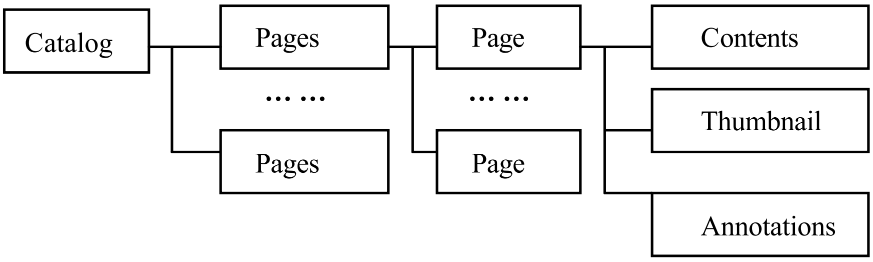
(3) Body : File body is made up of abundant PDF *Indirect Objects* . *Indirect Objects* constitute the specific contents of a PDF document as font , page-layout , table , image and so on . How to deal with *Indirect Objects* in file body is the focus of our work .

(4) Head : File head indicates the edition number of PDF criterion to which PDF documents conform . For example , % PDF-1.4 means that the document accords with PDF 1.4 criterion .

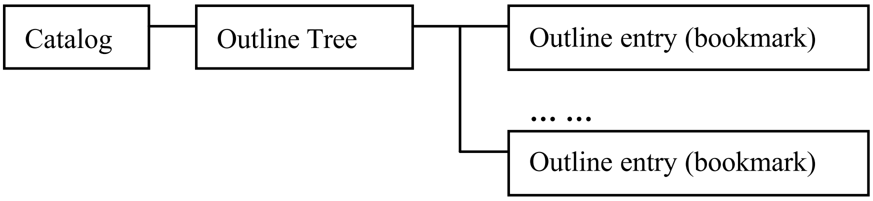
2.2.3 Logical Structure of PDF

The logical structure of PDF reflects the hierarchical relationship among the *Indirect Objects* of file body , which can be expressed by tree structure . The root node is just the PDF documents' root object *Catalog* . There are four subtrees below the root node .

(1) Pages Tree : All the page-objects are the leaf roots of this Pages tree . Each page contains the citation of content , note and breviary . Content Stream describes the text content of pages , so how to treat with content stream is crucial . Details see 2.3.1 .



(2) Bookmark : The Outline tree in PDF documents is also a tree hierarchical structure . Each node of the tree is a bookmark . The function of bookmark is to relate a bookmark name with a specific page position . According to bookmark names , application program can access the contents of PDF documents expediently . Details see 2.3.4 .



(3) Threads Tree : Organize the threads of article and the article blocks which are below the threads in terms of the tree structure .

(4) Name Tree: Associate a string with a page area .

2.3 Realization of Information Extraction from PDF Documents

2.3.1 Extract contents from content stream of each page , then decode

Firstly , application program accesses file trailer to obtain the address of cross table and root node *Catalog* . Secondly , application program visits the *Indirect Objects* of PDF documents by cross table . As a result , we are able to control the whole PDF documents .

(1) Look for PDF root object from file trail . The type of that object is *catalog* .

(2) Find page-tree root object through root node *catalog* . The type is *pages*.

(3) Search for page objects according to child node *pages*. The type is *page* .

(4) Access to the content of page object *page* . If application program fails to find the entry of *Contents* , leap over it and do nothing ; If it succeeds in finding the entry of *Contents* , jump to the next step .

(5) Attain all the *Contents* object number from the entry of *Contents* , then record the object number onto the array *Con_objNo[]* in sequence .

(6) According to the every object number in array *Con_objNo[ ]* , switch to the corresponding object position . Then extract the encoding name after *Filter* and put the contents between *stream* and *endstream* into the array *con_byte[ ]* .

(7) Assign the encoding means *Filter* of software package *iText* in Java to encode the contents stream of objects in the array *Con_objNo[ ]* .

(8) Using pointer , connect all the strings encoded from the objects in the array *Con_objNo[ ]* together . The outcome of this step is a string *TextStr* embracing the contents of the aimed page .

(9) Repeat the above steps until all the pages have been treated . Then application program also join all the encoded string together by pointer . Last , a intermediate file has been created .

Notes :

1 The organization of nodes in Page tree has the characteristic of Preorder Depth-First . We adopt Preorder Traversal arithmetic to read all the objects and attributes of pages , then write these contents into the intermediate file in turn . It turns out to be that the sequence of the visiting page nodes is the same as the actual page number of pages .

2 If the PDF documents is written in English , the encoded strings in intermediate file are just the original text . If the PDF document is a Chinese document , the encoded strings in intermediate file are the coding of Chinese characters . That is to say , it is necessary to transfer the codings to Chinese characters to get the original text .

2.3.2 Transfer the physical structure of PDF documents to logical structure

We can acquire the following important information from the intermediate file : (1) content : the text contents of each row of each page ; (2) position : the position of each row (x , y) ; (3) page : the page number that the treated row locates at : (4) font type : the font type that most text contents of the treated row adopts ; (5) font size : the font size that most text contents of the treated row adopts .

The intermediate file describes the physical structure of PDF documents , while have nothing information about semantics of PDF documents , so it is time for us to transfer the information from intermediate file to the logical structure of article that caters to people's reading habit .

(1) Analysis of the Page Space : The goal of this step is to transfer the intermediate file organized in physical rows to a file structured in logical rows . For articles which is typeset in single column , logical rows equate physical rows in a sense . But for articles typeset in many columns , application program must regulate rows according to columns. The core of this step is to distinguish strings located at different columns but the same row .

(2) Logical Transition : After the above treatments , the string chain-list arranged in the articles' physical sequence has been established . The work we are doing now is to convert the string chain – list to logical chain-list that people are familiar with . Our system uses clustering arithmetic to cluster the content of the same column together .

2.3.3 Extraction of the Exterior Information metadata

After the pretreatment illuminated above , we succeed in attaining the logical chain-list of articles . Now , it's time judged the exterior information metadata according to the DTD mentioned in 2.1 .

The regulations to pick up the first author are as follows : (1) The Y-coordinate of the string is the most close to that of the title which has been extracted . The way to judge an article's title will be explained in 2.3.4 . (2) The font size of string is less than that of title .

The rules to pick up the non –first authors are showed below : (1) The Y-coordinate of the string is equal to that of the first author ; (2) The font size and font type of the string is the same as that of the first author .

The rests are the basic information about authors ,such as college name , address , postal code .

2.3.4 Extraction of Interior Information metadata

Traverse the whole logical chain list to extract the semantic information .

(1) Title : Application program refers to the following regulations to draw out title : ① The Page of the string is the first one ; ② The Y-coordinate of the string is the biggest ; ③ The font size of the string is the largest . If a string accord with these terms at the same time , application program regards it as a part of the title .

(2) Section : Accordingly , the Outline Tree is a tree hierarchy structure , and each node is a bookmark . Application program makes use of Bookmark to extract information about sections . The workflow of this course is : Firstly , convert the depth of the Bookmark node in the Outline Tree to the hierarchy structure of sections in XML documents ; Secondly , map the text content of Bookmark to the element *Theme* of *Section* ; Thirdly , judge the paragraphs that a section contains by the specific position that the bookmark indicates .

(3) Para :

There are two regulations to judge paragraphs . On one side , if the space between two text rows is larger than the average space between , we can conclude that "These two text rows belong to different paragraphs " . On the other hand , if the x-coordinate of a text which lies in the front of a row is larger than that of the former row , we can perorate that " This row is the beginning of a new paragraph " .

An important information of a paragraph is the theme of the paragraph . There are two ways to denote a theme : The first one is the abstract form ; the second one is the keywords form . We adopt the keywords form to represent the theme of a paragraph in our system . Besides , our system employs Chinese Information Processing method to get the theme .

Chinses Words Divided Syncopation : It is the course of automatic text -recognition of computer and can be expressed by the function : a =F(b) . The character *b* means the Chinese character sequence ($b_1 b_2 \ldots b_n$) , while the character *a* is the Chinese word cluster ($a_1 a_2 \ldots a_m$). Our system uses positive Maximum Match arithmetic as F(b) . The arithmetic uses a word separation table and the principle"Long words first " to separate words . The basic idea of the positive Maximum Match arithmetic is as follows : Suppose that the number of the characters constituting words in the word separation table is i .Cut number i characters from the input Chinese character sequence as matching field first . Then look up the word separation table to find whether the matching field is in it . If the matching field is the same as some word in the word separation table , cut this matching field out and add it into array a[ ] ; If not , remove the last character from the matching field . Repeat the above steps until succeed in matching .

② Label part of speech : Use specific tools to label part of speech for all the vocabularies in array a[ ] .

③ select keywords : Judge all the nouns of a paragraph by the above result . Then calculate the word appearing frequency of the nouns . Refer to Shannon theory , the most meaningful words distinguishing a paragraph from others must be the words that the appearing frequency in the paragraph is adequately high , while that in other paragraphs are sufficiently low . So we consult a vector representation which is called TFIDF(Term Frequency Inverse Document Frequency) ,and define the following formula to compute the appearing frequency :

$x_i = freq\ (w_i)\ log\ (N/DF(w_i))$

Notes : *freq* (w_i) means the appearing time of w_i in a paragraph ; $DF(w_i)$ represents the number of paragraphs embodying the word w_i ; N is the total number of the targeted article' paragraphs . And then , select several nouns in possession of the most appearing frequency as the themes of this paragraph . Last , write them into the subelement *para_theme* of element *para*.

2.4 Building XML Document

After encoding , sections dividing , words separating and other treatments , our system will build a text-structured tree . On this basis , we can reach the final goal of transferring a PDF document to a valid XML documents using the designed DTD document as a template .

3 Conclusion

The article mainly dwells on the design frame of the PDF documents information extraction system and the correlative techniques of how to transfer a PDF document to a XML document. Users are able to do further operation to the XML documents so as to the efficiency of document classification and information retrieval will be improved .

References:

[1] Adobe Systems Inc.，PDF Reference ： Adobe Portable Document Format version 1.4_[M].3nd ，2001

[2] Extensible Markup Language 1.0Second Edition, *http://www.w3.org/TR/REC-xml,2000-10*

[3] http://www.docbook.org/xml/simple/1.1CR2/

[4] Yang Daoliang ,.The Design and Implementation of an Object- Oriented Chinese PDF R eader. Computer Application，1999，19（6）

[5] Introduction to XML ,Java, databases and the web Nazmul Idris 1999/06/24 http://www.developerlife.com

[6] Norbert Fuhr. XML Information Retrieval and Information Extraction . *http://ls6-www.informatik.uni-dortmund.debibfulltext/ir/Fuhr:02a.pd,2002*

[7]Xu Jinfeng , Basic Tutorial for Chinese Information Processing . Pekin University Press , 2002

[8] LiHui,Shi Zhongzhi . Improving the Performance of the Text Classif ier Based on Support

VectorMachine Using the Common Sense in Text Domain . Journal Of Chinese Information Processing , 2002 Vol.16 No.2

[9] Ekkuitte Rusty Harold. XML Bible [M]. China WaterPower Press,2001