ATLANTIS PRESS
Advances in Engineering Research, volume 141
5th International Conference on Mechatronics, Materials, Chemistry and Computer Engineering (ICMMCCE 2017)

Research and Application of Word Format Checking Technology based on Java and XML

Lu Han[1, a], Kun Liu[1, 2, b, *] and Jinmin Jiang[1, c]

[1]School of Information Science and Engineering, University of Jinan, Jinan 250022, China

[2]Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan 250022, China

[a]869250124@qq.om, [b]ise_liuk@ujn.edu.cn, [c]460428795@qq.com

Abstract: Through researching the internal organizational structure of Word 2007 document, the significant component documents are extracted, including document.xml, style.xml, header.xml and footer.xml. The compositions structure and the interaction of the common document formats in each XML file are analyzed. Accordingly on that basis, the Java and XML technologies are adopted to extract the format mark and the element property of the documents as to further anatomize the document format. Such combined technology shall be adopted to check the format of degree papers as to further boost the information progress of the education management for college and university.

Keywords: Format Checking; Word; Java; OOXML

1. Introduction

The XML format adopted after the series of Microsoft Office 2007 has been perceived as the format. Currently the XML format has been deemed as the standard international document format, greatly boosting the new document processing method of the Microsoft Office. The Open Packaging Conventions are recognized as the basic convention abided by the base of Office Open XML (OOXML) to describe the organization, packaging and presenting of the document content. The Office Open XML (OOXML) adopts the XML as the document format, and applies ZIP technology as the RAR storage format. All these formats can be adopted cross the platform. Additionally, through these formats, the reliable and conformable document generation and data exchange among the platforms can be realized.

Given that the degree papers of all colleges and universities are diversified and distinctive, the automatic checking and matching system of the paper format is currently facing the inaccurate format orientation, the incomplete checking items, etc. For this reason, the computer system adoption to check the paper format is desiderated by the students, teachers and the universities. Given the information mentioned above, the research of the Java and XML-based word document format checking technology and the application of such technology in the paper format checking of colleges and university are attached great practical significance and the feasibility.

2. WordprocessingML Document Analysis

Office Open XML (OOXML) is the international document format standard formulated by Microsoft. It is literally the standard specification of the electric documents compressed in ZIP form based on the XML technology. The document formats, including file, spreadsheet and memorandum. WordProcessingML is the markup sentence adopted in the OOXML to generate and establish the word document, as the component of the OOXML. The basic document structure of WordProcessingML documents is composed by the elements <document>and <body>, attaching one over multipleblock-level referring to the paragraph, like <p>. One or multiple <r> elements are

contained in the paragraph. <r> refers to the consecutive text. It is the text area containing a group of common properties (including format setting). A paragraph of consecutive text shall contain one or multiple<r> elements. The <r> element shall contain a series of text.

3. WordProcessingML Document Parsing and Processing Level

The checking of paper format is deemed as the process to process and compare with the WordProcessingML documents through adopting the Java-related technologies. Such process shall contain three levels, viz. the parsing level, storage level and processing level. These three levels are adopted to parse, store and compare with the formats of WordProcessingML document. The three-level model is displayed as in Fig. 1:

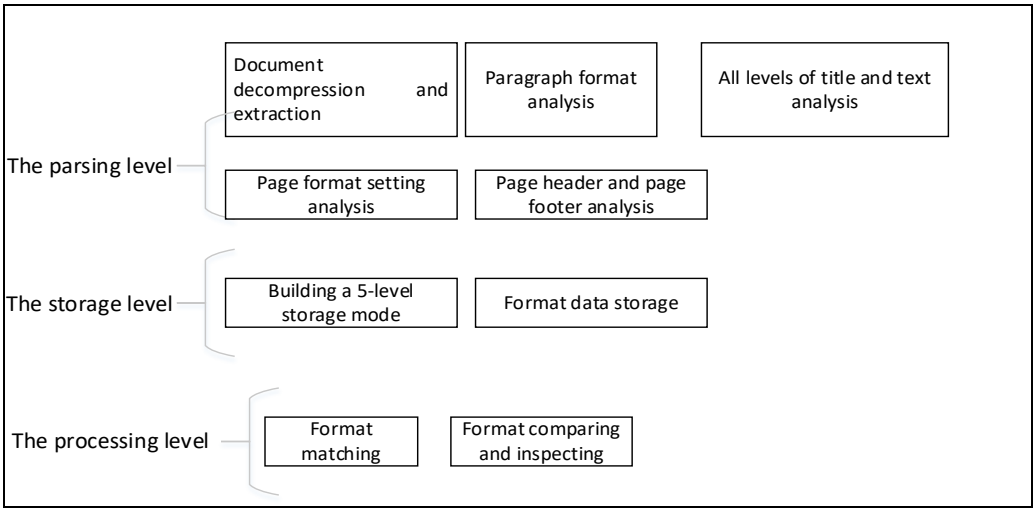

Fig.1 Three-level Model

3.1 WordProcessingML Parsing Level Design

WordProcessingML parsing level is mainly applied to parse the format of document.xml document. Through adopting the Dom4j technology, the content is firstly markled and secondly extracted.

Mainly two functions exist in the parsing level. The first one is the paper model format set by the parsing user to generate the corresponding model. The second function is to parse the formats of all models of papers to be checked, and further to carry out the format comparison in the processing level.

Document Decompression and Extraction. No matterwhat the content is to be parsed, whether the paragraph format, page setting parsing, page header or footer parsing. The first thing referred to be done is to decompress the word document based on the OOXML standard and to further acquire the corresponding document through the byte stream or the character stream. Additionally the file reading and format mark shall be parsed. As the word document based on the standard of OOXML is decompressed, multiple XML documents and multiple folders are contained. For this reason, to parse the format of the document, the first thing required to be done is to parse the XML document through adopting the Java-related technology. The Dom4J is flexible and variable in structure. It can be adopted conveniently. The Dom4J can be adopted hereby parse the XML document.

Parsing Paragraph Format. The basic format of the paragraph is composed by the paragraph format and character style format. The paragraph format is composed by the alignment, indentation for the first line, line spacing among the paragraph and other contents. The character style includes the Chinese character style and the English character style. The format content is mainly composed by the character category, character size, whether be bold or not and other elements. With the understanding of the basic structure composition of the paragraph, the parsing of the paragraph

format can be carried out according to this.

First and foremost, the paragraph mark in the document.xml document is extracted as <w:p>. The paragraph mark contains the <w:pPr>mark and the <w:r>mark. The markled <w:pPr> refers to the paragraph format. If the <w:pStyle>mark is contained in the <w:pPr>, the sytle.xml is read. The <w:spacing>, <w:indent> and <w:jc>marks are extracted through the property of w:val. The indentation for the first line, line spacing among the paragraph can be analyzed through markling the <w:spacing>, <w:indent> and <w:jc>. The character content and the character format are contained in the <w:r>mark. <w:t>refers to the character content. <w:rPr>refers to the character format. If the <w:pStyle>mark is contained in the <w:rPr>, this document is read as sytle.xml. The <w:rFonts>, <w:sz> and <w:b>marks are extracted through the property of w:val. The character category, character size, whether the character is bold are anatomized through the <w:rFonts>, <w:sz> and <w:b>marks.

Analysis of Page Setting Format. The specific format structure of the page setting is composed by the page size and format and the page margin format. The page size format is mainly composed by the page height and the page width. The page margin content is mainly composed by the top margin, bottom margin, right margin and left margin. In accordance with the basic format composition of the page setting, the format of the page setting can be parsed.

First and foremost, the <w:sectPr> mark is extracted from the document.xml document. The <w:pgSz> and <w:pgMar> marks are contained in the <w:sectPr>. The <w:pgSz>mark refers to the page size setting, containing the property w:w and property w:h, viz. the page width and page height. <w:pgMar> refers to the page margin setting, including the properties of w:top, w:right, w:bottom and w:left，viz. the top margin, right margin, bottom margin and left margin.

Analysis of Page Header and Footer Format. First and foremost, the <w:sectPr> mark is extracted from the document.xml. The <w:footerReference> and <w:headerReference> marks are contained in the <w:sectPr>. Through marking the property r:id in the marking process to the reading of corresponding format in the footer.xml or header.xml, the specific operation content after extracting the format is similar to the paragraph format checking, and compared with the page header or footer format in the database.

3.2 WordprocessingML Storage Level Design

The WordprocessingML storage level is mainly used to store and analyze the structure format data of the storage analysis level. Through studying the marking sentence of WordProcessingML, the five-level storage model, including template, model, format, element and property, is established. The data processed of the analysis level is stored in accordance with the level. The five-level storage model is illustrated in the Fig. 2.

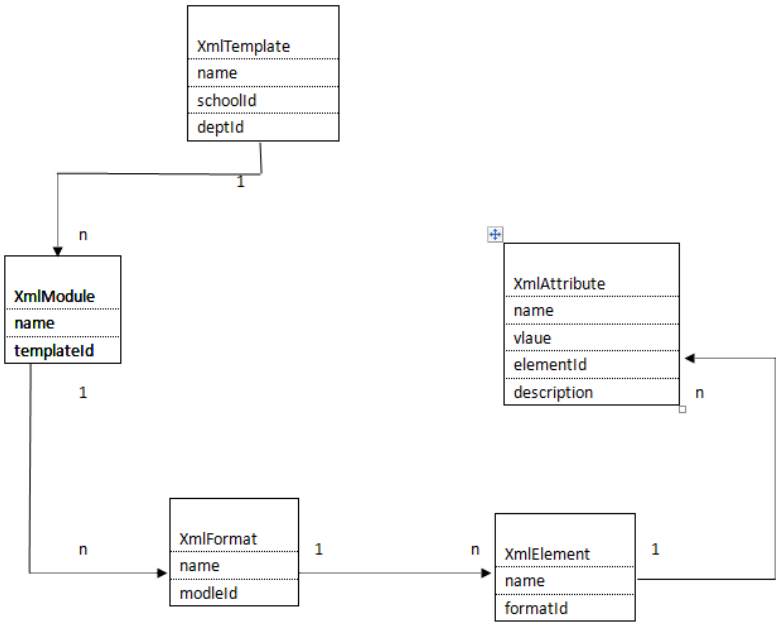

Fig.2 Five-Level Storage Model

Under such five-level storage structure, multiple models are contained in one template, multiple formats are contained in one template, multiple elements are contained in one format, and multiple properties are contained in one element.

The template name, model name and the format name are set artificially while we are setting the paper model. The elements and the properties are extracted and stored from the xml documents. For instance, the line spacing format of a certain paragraph is marked as <w:spacing w:line="600" w:lineRule="auto"/>, where 600 acquired through multiplying the line spacing by 240 (line spacing standard). For this reason, the line space shall be 2.5 times. w:spacing shall be preserved through adopting XmlElement. The name is set as w:spacing. Both the w:line and w:lineRule are the properties of the w:spacing element. The 600 from the w:line="600" is the value of w:lineproperty which shall be stored through adopting the XmlAttribute. The names shall be set as w:line, and the value is 600, described as the line spacing of the content. The conversion of the property and the description is carried out through one conversion tool. The property shall be described through adopting the sentences that can be understood by the public.

3.3 WordprocessingML Processing Level Design

The WordprocessingML processing level has two major functions. One is that this level can be used to compare and check the formats between paper and model. The format data of the paper to be checked after the checking in the analysis level and the format data of the stored template document are compared with each other as to generate the checking report. Secondly, this level is applied to check the blurred picture, picture, graphical representation and the inter-page of the papers.

Checking the Format through Comparison. The comparative checking of the document format is required to firstly preprocess the document, and to secondly generate the collection of format to be checked through extracting and analyzing the format in different models. Lastly each model shall be checked in the format respectively. While checking the format, whether the element to be checked is banked or not shall be confirmed firstly. For instance, the alignment mark <w:jc val="center"> is contained in the models, whereas such mark is not contained in the paragraph to be checked. If this element to be checked is blanked, the null value processing shall be carried out. If the element to be checked exists, the follow-up matching format element operation shall be initiated. Accordingly the element to be checked shall be matched with the format element of the same model in the template

as to find out the corresponding format element of the template. The character size property mark can be manifested in two approaches, viz. the w:sz and w:szCs which shall be special processed. The format properties shall be compared in accordance with element name, and the compared result shall be transformed. The w:beforeLines shall be transformed as the line spacing in front of the paragraph, the w:lineshall be transformed as the line space, and w:val="center" shall be transformed as the align center. If some elements are blanked, these elements shall be carried out the default processes.

Blurry Degree Checking of Picture.First and foremost, the picture resource of the document is acquired. The jython technology is adopted, and the python script is called. Accordingly the basic blurry checking is realized. The Laplace is adopted to carry out the convolution operation for the grey level of the picture, and accordingly the variance shall be calculated (viz. the square of the standard deviation). If the variance of a certain picture is lower than the threshold defined previously (100), such picture shall be deemed as the blurred picture. Otherwise the picture is deemed as the clear picture.

4. Document Format Checking Result and Analysis

The word document format checking mainly contains the format checking of essential paper models, including Chinese title, the content of Chinese abstract, English abstract content, Chinese title, English abstract keyword title, keyword title keyword content in Chinese and English, English keywords content, primary title, secondary title, third level title and the icon and graphical representation. The elements to be checked include the character style, character size, alignment, paragraph line space, page header and footer, page setting and other common formats. To examine the practical effect of the platform checking the paper's format, a paper is carried out the artificial checking and the checking via the platform program. The results are compared. Partial of the checking results are presented in Fig. 3.

Error place	Error content
Chinese abstract and the second paragraph of text	Typeface of [00XKL] is error. Song type is wrong and Times New Roman type is right.
Chinese keywords	Typeface of the punctuation [:] is error. Black type is wrong and Song type is right
Englishabstract and the second paragraph of text	Chinese characters should not exist in English abstract.
English keywords	The punctuation of [:] should not in Chinese type;
The first paragraph of text	The value of line spacing is wrong and it should be 1.5 but 1.25. The typeface of preface is error. Black type is wrong and Song type is right. The size of typeface is error. It should not be less No.3 but less No.4
The title of second level and the second paragraph of summary	The value of line spacing is wrong and it should be 1.5 but 1.25. The size of"word" typeface is error. It should not be less No.3 but less. The typeface of "text format" is error. Black type is wrong and Song type is right. The size of typeface is error. It should not be less No.3 but less No.4

Fig. 3 Checking Result

Compared with manual detection, the program detection is more accurate and comprehensive, and

there will be no omission existing in the process of checking. In the meantime, for a 450-page paper, the full text can be retrieved in a few seconds, and the error format report and correction method can be generated at the same time.If you use manual retrieval, it will take at least a few dozen minutes, and the accuracy is not high, thus greatly improving your efficiency.

5. Conclusion

This paper analyzes the Microsoft Office 2007 and the OOXML format adopted in the follow-up stage, anatomizes the organizational structure of the word document and the function and significance of the relevant component documents. The composition and structure of the WordprocessingML document is delved into. Additionally, the WordprocessingML marking adopted in the OOXML is introduced, and the WordprocessingML elements and the marks commonly adopted in the documents are described.

Through analyzing the analysis level, storage level and the processing level of the WordprocessingML document, this paper highlights the introduction of the core technology adopted to check the format of the document. The document decompression, paragraph format analysis, title and main body analysis, analysis of page footer and header, page configuration and other core parts are delved into. Eventually, through artificially checking and checking via program, the research result is practically verified and tested.

This paper analyzes the Java and XML-based word document checking technology and applies such technology in the checking of paper format for colleges and universities. Accordingly the efficiency of the paper format checking can be effectively elevated. Certainly this system requires to be further optimized, including the efficiency of paper format checking to be further optimized, whether can carry out the preloading or pre-checking, and the improvement of user's experience. In the meantime, a relatively complete score planning is required to be proposed as to present the valued checking score for the paper formatchecking result, which can reflect the paper format matching condition to some extent. These aspects shall be further researched.

Acknowledgment

This work was supported by Shandong Provincial Key R&D Program (2016ZDJS01A12).

References

[1] Dong W. A Web Service Based Management System Of Degree Theses: Research And Implementation [J]. Web Services Soap Management System Degree Theses, 2014:781-786.
[2] Surhone L M, Timpledon M T, Marseken S F, et al. Standardization of Office Open XML [M]. Montana:Betascript Publishing,2010.
[3] Lagadec P. OpenDocument and Open XML security (OpenOffice.org and MS Office 2007) [J]. Journal of Computer Virology and Hacking Techniques, 2008, 4(2):115-125.
[4] Mirakhorli M. Software architecture reconstruction: Why? What? How? [C]. IEEE International Conference on Software Analysis, Evolution and Reengineering,2015.
[5] Yenig, N, H, Yilmaz C, et al. Advances in test generation for testing software and systems [J]. International Journal on Software Tools for Technology Transfer, 2016, 18(3):245-249.
[6] Levina E, Mustafina G, Nigmetzyanova V, et al. Improving the Information System of University Management [J]. Review of European Studies, 2015, 7(1):109-116.