

The Design of Automatic Scoring System for Document Operational Test Question of MS-Office

Hao Li

College of Computer Science
Inner Mongolia University
Hohhot, China
gongyangcheng@163.com

Guan Qingguo

Strategy and Finance Department
Inner Mongolia postal Corporation
Hohhot, China
haoli@imu.edu.cn

Abstract—Automatic scoring system is helpful to efficiently scoring the document operational test question of MS-Office. The existed scoring methods have done a lot of valuable works. But these methods record the operation track and lead to misjudge when modifying operations. In this paper, we propose a new method for this problem. The basic idea of this method is using Office Open XML file. First, extract the corresponding XML file from MS Office 2007 document of standard answer and MS Office 2007 document of student's answer. Next, score the student's MS Office 2007 document by comparing the similarity of two XML documents. Using Windows as our experiment platform, we did a lot of tests. The experimental results show that the error evaluation rate of this system is less than 3% compared to the manual scoring.

Keywords—MS Office 2007; Office Open XML; document automatic scoring;

I. INTRODUCTION

Operation and application of MS Office(Word、Excel and PowerPoint) software occupies an important position in the computer application basis course of tertiary institutions, and the examination formats and test paper grading methods of this part have also been a hot topic among teachers. Nowadays, generally accepted evaluation mode is on-line test. Because computer application basis course is the public basis course oriented to various majors of school, so the number of course selection is more and the workload of test paper grading is heavy, while using computer program to automatically score the student's answer will, to a large extent, reduce the workload of teachers and improve the accuracy and justice of paper marking [1].

Currently, most automatic scoring system for document operational test question of MS-Office is basically built on the basis of MS Office 2000 or MS Office 2003, and the technology adopted mainly combine VBA macro language and OLE technology to write object-oriented program to automatically score MS Office operation question [2]. Through study and summary of related technology which can automatically marking the Office document, we find that we mainly use the following three techniques in practice, namely: ① Automatic marking after converting Office documents to RTF documents; ② Using OLE(Object Linking and Embedding) automation application program to realize automatic marking; ③ using VBA (Visual Basic for Applications) technology to realize automatic marking.

Among which, the third technique is currently the most adopted technology in research and implementation of automatic marking system for Office document [3,4].

This paper proposes a new method in designing automatic scoring system for MS Office operation question, namely on the basis of analyzing Office Open XML document format which is adopted by MS Office2007, comparing the similarity of the content of student's document with the content of standard document to implement automatic marking for MS Office document[5].

II. THE STRUCTURE OF THE OFFICE OPEN XML FILE FORMATS

Office Open XML file format is a ZIP file format. Office Open XML format uses some XML reference frameworks and a ZIP container. Each file is made up of some parts set, which defines the document. Document parts are stored in a container file, or stored in a ZIP format package based on the industry-standard. Many parts are used to describe application data, metadata, and user-defined XML data, which are all stored in the container file. Other non-XML parts can also be included in the container package, including the binary file representing pictures in document or embedded in OLE object. In addition, there are relational parts that specify the relationship between parts. Parts make up the contents of the file, while relationship describes how the content is organized together[6]. Figure 1 shows the Office Open XML file format container of 2007 version:

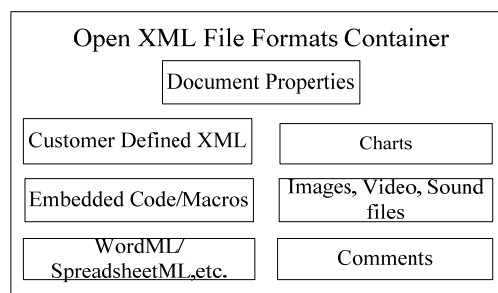


Figure 1. Office Open XML File Formats Container of MS Office2007 Version

III. AUTOMATIC SCORING ALGORITHM

From the structure of MS Office2007 file formats, we know that MS Office2007 document consists of many

logical parts, and each part is made up of separate XML files. Take MS Word2007 as an example, main logical units included by MS Word2007 document are document contents, pictures, graphics, page headers and page footers, comments, format definition, list definition and so on, and each part is composed of XML files, and a series of simple tags in the XML file are combined together to represent the corresponding specific part in the document. Thus by processing the corresponding XML file, we can get the corresponding information in Word2007 document. So we can construct the following algorithm to implement automatic scoring: extract the related node information from XML file of standard answer document and compute node weight; the obtained node data information and the corresponding node weight will be imported into database; the XML file of student's answer documents will be processed according to the algorithm processing the standard answer, and then the obtained information and the information in the database will be carried on matching and scoring [7]. The automatic scoring principle of automatic scoring system for MS Office2007 is shown in Figure 2.

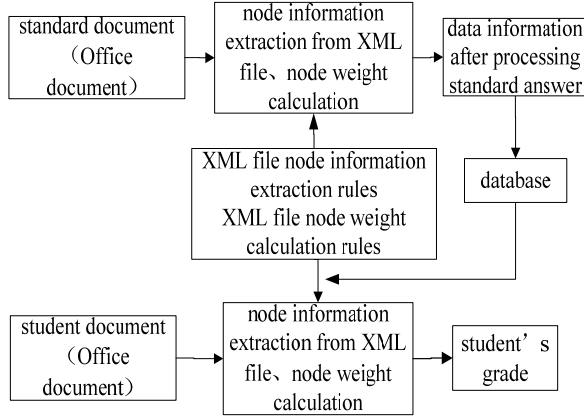


Figure 2. MS Office2007 automatic scoring system's automatic scoring principle

Next, take document.xml file in MS Word2007 document for an example in order to discuss the automatic scoring algorithm in the automatic scoring system for operation question of MS Word2007.

A. Introduction to the Document.xml File Structure

Document.xml file is MS Word2007 document's main document part, which includes the document's primary content (except page headers, page footers, footnotes etc). Document.xml file contains one root node <w:document>, and root node only contains one first-level child node <w:body>. There are <w:p> child node representing paragraph, <w:tbl> child node representing table and <w:sectPr> child node representing page setup under <w:body> node. And the order of arrangement of each node in document.xml file is the same as the display content order of the contents represented by this node in the MS Word2007 document page.

B. Automatic Scoring Algorithm

From the analysis of document.xml document structure, we can view document.xml document as a collection of nodes. Mean while, taking into account the characteristic of the contents of student's answer document consisting of multiple segments, we can use scoring method according to segment when scoring.

There are n first-level child nodes below the node <w:body> in the document.xml document, viewing it as a set containing n elements (A_i). We use Sample to represent the document.xml document of standard answer document, then Sample document can be expressed as Sample={ $A_1, A_2, A_3, \dots, A_i, \dots, A_{n-1}, A_n$ }. Among which, A_i contains four properties: name, value, length and boolean. A_{iname} refers to the name of first-level child node represented by A_i , A_{ivalue} refers to the string composed of properties of all child nodes which are contained by first-level child node represented by A_i , $A_{ilength}$ refers to length of string represented by A_{ivalue} , $A_{iboollean}$ refers to whether A_i has carried on similarity calculation.

Suppose that w_i represents the weight of each element A_i , then its weight is:

$$w_i = \frac{A_{ilength}}{\sum_{i=1}^n A_{ilength}} \quad (1)$$

Suppose that the score of Sample document is C, and the corresponding score of each element A_i is represented by c_i , then $c_i = C \times w_i$.

We use Draft to represent the document.xml document of student's answer, viewing Draft document as a set containing m elements(B_j), then Draft document can be expressed as Draft= { $B_1, B_2, B_3, \dots, B_j, \dots, B_{m-1}, B_m$ }. Among which, B_j contains two properties: name and value. B_{jname} refers to the name of first-level child node represented by B_j , B_{jvalue} refers to the string composed of properties of all child nodes which are contained by first-level child node represented by B_j .

Match element B_j in Draft document with element A_i in Sample document, and suppose that the similarity of B_j with respect to A_i is L_i , then L_i can be calculated by the following algorithm:

```

Initialization: for i=1 to n, let  $L_i = 0$ ,  $A_{iboollean} = \text{false}$ 
For j=1 to m
  For i=1 to n
    If  $A_{iname} = B_{jname}$ , and  $A_{ivalue}$  is not NULL,  $B_{jvalue}$  is not NULL,  $A_{iboollean} = \text{false}$ , then
       $L_i = \text{sim}(A_{ivalue}, B_{jvalue})$ 
       $A_{iboollean} = \text{true}$ 
    Exit i loop
  Next i
Next j
Among which,  $L_i = \text{sim}(A_{ivalue}, B_{jvalue})$  is used to calculate the similarity between two strings.
The similarity of two string strA and strB can be defined as:
  
```

$$(2) \quad \text{sim}(\text{strA}, \text{strB}) = \begin{cases} \frac{\text{strAlength} - \text{LD}(\text{strA}, \text{strB})}{\text{strAlength}}, & \text{strAlength} - \text{LD}(\text{strA}, \text{strB}) \geq 0 \\ \frac{\text{strBlength} - \text{LD}(\text{strA}, \text{strB})}{\text{strBlength}}, & \text{strAlength} - \text{LD}(\text{strA}, \text{strB}) < 0 \end{cases}$$

Among which, strA.length represents the length of string strA, strB.length represents the length of string strB, LD (strA, strB) represents the number of different characters of string strB against string strA, obviously, the value of sim(strA, strB) is between 0 and 1. String strA and strB is completely different when the value of sim(strA, strB) is zero, and string strA and strB is exactly the same when the value of sim(strA, strB) is one. We adopt edit distance algorithm (Levenshtein Distance) when calculating LD(strA, strB)[8,9].

Calculating Draft document according to L_i , the obtained score sum is:

$$sum = \sum_{i=1}^n L_i \cdot c_i \quad (3)$$

We realize scoring a XML document part through the above algorithm, and a MS Office2007 document usually consists of multiple XML document parts, so we can calculate score of MS Office 2007 document through the following formula, that is:

$$G = \sum_{i=1}^m sum_k \cdot P_k \quad (4)$$

Among which, m represents the number of XML document parts needed to score contained by MS Office 2007 document, P_k represents the k XML document part needed to score in Office 2007 document, sum_k represents the score of the k XML document part, G represents the final score of Office 2007 document.

IV. THE REALIZATION METHOD OF AUTOMATIC SCORING SYSTEM

The main function of this system is to realize automatic scoring for document operational questions of MS Office2007. The function mainly contains three sub-functions, namely, automatic scoring for document operational test of Word2007, automatic scoring for document operational test of Excel2007 and automatic scoring for document operational test of PowerPoint2007, which can be divided into three sub modules to implement. Meanwhile, in order to meet system compatible with Office2003 test paper documents, a function module of converting Office2003 test paper document into Office2007 test paper document is added to the system. Each function module is designed to meet the demands of high cohesion and low coupling as far as possible, so that system in subsequent development has better reusability, maintainability, and scalability.

The design of operation flow of this system uses modular flow design, that is, the operation flow of automatic marking contains module one—converting Office2003 documents into Office2007 documents module—and module two—automatic marking module for test paper document of Office2007. Steps of specific operation flow are as follows:

Step 1: determine whether the document submitted by the student is Office2007 document. If it is not Office2007 document, then proceed to step 2; otherwise, proceed to step 3.

Step 2: Select the corresponding conversion sub-modules from module one, and select the Office2003 documents needed to transform on the local computer, then the selected document will be imported into sub-module, the system will automatically convert the imported document into Office2007 document. The specific operation flow is shown in Figure 3 module one:

Step 3: Select the corresponding marking sub-module from module two, then select the corresponding standard answer document on the local computer and import them into system, next, select and import the student's answer document, after the student's answer document preprocessing has been completed, click the marking button in the system, and the preprocessed test paper document will be carried on marking and scoring, and finally export the student's grade, thus one paper marking has been completed. The specific operation flow is shown in Figure 3 module two:

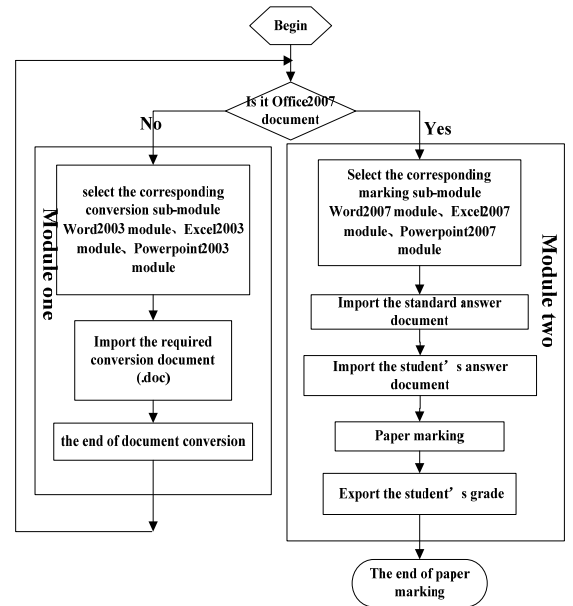


Figure 3. Automatic Marking Operational Processes

V. TEST AND CONCLUSION

This system selects 2485 word document, 2097 excel document and 1109 PowerPoint document, which are made by students, to carry on testing. The document is divided into Word, Excel, and PowerPoint three types when testing, and each type is divided into 10 sample groups. Using this system to automatically score the documents of sample groups, then comparing the automatic scoring results with manual scoring results, the measured average mistake rate for different type of documents are as follows: Word document is less than 2.9%, Excel document is less than 1.9% and PowerPoint document is less than 1.8%.

Here, we extracted 60% of the test samples using simple random sampling method. And the experimental results of comparing automatic scoring with manual scoring are shown in Figure 4, Figure 5 and Figure 6.

group	Test Paper Number	Highest Score (automatic /manual)	Lowest Score (automatic /manual)	mistake (%)	Average mistake (%)
1	134	94.05/95	10.15/12	2.5885	2.87744
2	155	99.35/99	22.02/19	0.7473	
3	172	99.16/99	17.46/16	3.9606	
4	137	94.10/92	26.72/26	3.4444	
5	119	98.93/99	38.38/39	3.1889	
6	118	90.57/92	24.18/26	3.2051	
7	163	93.52/92	14.45/15	1.5848	
8	154	98.13/99	30.19/31	2.9112	
9	172	90.09/91	28.46/30	3.2047	
10	167	91.38/93	35.02/35	3.9389	

Figure 4. Word group's test results

group	Test Paper Number	Highest Score (automatic /manual)	Lowest Score (automatic /manual)	mistake (%)	Average mistake (%)
1	152	97.09/98	24.56/25	0.6151	1.80028
2	140	94.28/94	20.18/22	1.8994	
3	137	93.04/93	34.05/35	3.1894	
4	145	91.89/93	38.13/40	1.3115	
5	125	91.93/91	28.21/30	0.6475	
6	109	96.82/95	38.93/40	3.9214	
7	97	93.02/93	21.76/20	1.2749	
8	112	95.41/95	16.03/15	2.4821	
9	116	91.50/91	24.44/25	0.6728	
10	125	96.97/96	26.15/25	1.9887	

Figure 5. Excel group's test results

group	Test Paper Number	Highest Score (automatic /manual)	Lowest Score (automatic /manual)	mistake (%)	Average mistake (%)
1	66	95.46/98	31.50/32	1.2711	1.75064
2	73	94.44/95	33.78/34	2.7889	
3	63	96.94/98	23.79/24	1.8333	
4	69	96.21/96	14.65/15	3.3965	
5	73	97.94/98	24.28/25	1.2084	
6	61	99.56/99	12.02/12	2.4865	
7	65	95.22/95	18.38/18	0.56	
8	64	98.80/99	21.93/22	0.5854	
9	67	91.72/91	35.41/35	1.6257	

Figure 6. PowerPoint group's test results

From the test results can be seen, the design of three automatic scoring modules is reasonable, and the mistake rate of automatic scoring result is low.

When designing the system, we consider the evaluation of document's final result, and we don't need to consider the contents of test paper. Automatic scoring system and

questions are separate, so teachers don't need to choose questions according to paper marking system in the process of choosing question. This system mainly realize automatic marking for MS Office2003 and MS Office2007 documents, the intelligent marking technique used by this system in the automatic marking aspect for document operational test question of MS Office can better meet the requirements of automatic marking to answer document submitted by students in the process of computer application basis course examination, so teachers are liberated from the heavy manual marking work.

The research results in this paper have been used in the examination of the Inner Mongolia University, Fundamentals of Computer Application course.

ACKNOWLEDGMENT

This work was supported by Inner Mongolia University of Computer Application Foundation Excellent course construction and Natural Science Foundation Project of Inner Mongolia, Grant No.2013MS0902.

REFERENCES

- [1] Xiaobing Shao, Huifen Liao. Implementing Auto-judge of Office Operation [J]. Computer and Modernization, 2005,12.pp 135-136.
- [2] Wei Yan, Xinpeng Yin. The Applications of VBA in Designing Computer Basic Course Examination System [J]. Journal of Qiqihar University, 2006,5.pp 59-61.
- [3] Mingqiao Zhu, Yue Chen. Design and Implementation of an Automatic Examination Paper Marking System [J]. Computer Engineering and Science, 1999,21.pp 66-70.
- [4] Decai Zong. Design and Implementation of Automatic Scoring System for Operational Test Question[J]. Computer Engineering and Design, 2010,31.pp 1156-1160.
- [5] Zhixie Qu, Yanheng Zheng, Dong Han. Design and Realization of a XML Automatic Marking System[J]. Computer Engineering, 2009,9.pp 189-191.
- [6] <http://www.microsoft.com/china/msdn/library/office/office/OfficeOpenXMLFormats.aspx?mfr=true>
- [7] Qi Du, Zheng Gong. Algorithm Realization of the Automatic Rating Based on the String Similarity[C].2011 National Conference on Electronic Information Technology and Application, 2011,12.pp 1000-1004.
- [8] Xingchun Diao, Mingchao Tan, Jianjun Cao. A String Similarity Calculation Method with a Variety of Edit Distance[J]. Computer Utility Research, 2010,27(12).pp 4523-4525
- [9] Zheng Yang, Lijuan Duan, Yingxun Lai. Discovery Technique of Network Short Text Public Opinion Hot Spot Based on String Similarity Clustering[J]. Journal of Beijing University of Technology, 2010,36(5).pp 669-673.