PROCEEDINGS OF SCIENCE

Statistical Methods to Evaluate Important Degrees of Document Features

Xia Hou[1 2]
Computer School, Beijing Information Science and Technology University;
Beijing Key Laboratory of Internet Culture and Digital Dissemination Research,Beijing, 100101, China
E-mail: houxia@bistu.edu.cn

Ning Li
Computer School, Beijing Information Science and Technology University;
Beijing Key Laboratory of Internet Culture and Digital Dissemination Research,Beijing, 100101,China
E-mail: ningli.ok@163.com

Hong-bo Yang[2]
Experimental Teaching Center of Electronics information and Control, Beijing Information Science and Technology University ,Beijing, 100101 China
E-mail: anonbo@bistu.edu.cn

Documents are often analyzed based on the features. Evaluation of the degree of importance of such features is the basic work of document analysis. Now, the work is primarily done by the human beings although the workload is excessively large and the results are subjective. Three statistical methods are hereby proposed on the basis of the actual usage of features in a large number of documents. The statistical values and those from experts are contrasted and analyzed. And then a selection based on voting thus is given. Some results from 500 document samples are also given.

CENet2015
12-13 September 2015
Shanghai, China

[1]Speaker
[2]Correspongding Author
[2] This work was supported by the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions (CIT&TCD201504056, CIT&TCD201304115) and the Project of Construction of Innovative Teams and Teacher Career Development for Universities and Colleges under Beijing Municipality (IDHT20130519).

PoS(CENet2015)078

1. Introduction

Office document is one kind of the most important information carriers. Till now there are more than one office document format, such as Open Document Format (ODF) [1], Office Open XML Format (OOXML) [2] and Uniform Office Format (UOF) [3]. Multi-formats give users more choices; at meanwhile, the interoperability problem occurs because such formats have different XML structures and semantics.

In order to improve the interoperability, some researchers focuses on analyzing these document formats and their relationships. It is concluded that the main corresponding relations and differences between OOXML and ODF in two papers [4-5]. This method based on features is common and important [6-8]. A document is described as a set of features, for example, a word document includes features of sections, paragraphs and tables, etc.

In fact, different document instances are composed of different feature sets so that the interoperability of them is different and even has great disparity. Paper [9] designed and implemented a system which can extract the set of features from a document instance and compute the interoperability of the instance. In the course of computing the interoperability, the factor of degree of importance (DI) of features is a kind of basic and important parameters. Until now, the values of features' DI are evaluated mainly by experts in the field of document processing. There are more than one thousand features in each document format, whatever ODF, OOXML or UOF; therefore, the workload of evaluation is too huge. Moreover, values from a certain expert rely on the personal experience so that experts maybe give different evaluation results for a same feature. Additional, DI values from experts are mainly depended on the view of format so that the values maybe not fit or reasonable for a certain document instance.

In fact, experts often consider how often a feature is used in practice when they decide the features' DI; therefore, the usage frequency is the main factor to evaluate DI. Some statistical methods to evaluate features' DI are proposed in this paper, and a voting method is also proposed to decide the final value of DI according to the expert value and the several statistical values. We have implemented a prototype system to evaluate DI automatically based on a large number of sample documents.

This paper is organized as follows. Part 2 introduces the basic work including a document model (Feature Data Model) and the concept of degree of importance which are researched in our previous work. Part 3 presents three kinds of statistical methods to evaluate the values of DI. Part 4 explains the design of a prototype system which can automatically compute DI values based on the statistical methods presented in Part 3. Some data from experiments are given and the differences among the three statistical methods are analyzed based on experiment data; in addition, some voting Rules are presented to decide how to make a choice according to the values of the three statistical methods and experts. We conclude our work in the last part.

2. Basic Work

In our previous work, we implemented a measuring system to compute document interoperability. More details can be seen in two papers [9-10]. In our work, a Feature Data Model (FDM) is constructed based on multi-formats, especially OOXML, ODF and UOF. In documents, a main kind of relations among features is a whole-part. For example, as a section contains one or more paragraphs and tables, a paragraph can contain one or more runs, the structure of FDM is hierarchical typically and called "Feature Tree". In FDM, features are abstract according to the public semantic among different document formats; moreover, FDM includes some interoperability parameters like Degree of translation (DT), DI and the Mapping Information (MI). All of these parameters and MI are assigned to the leaves in the feature tree. For word-processing documents, the structure of FDM is illustrated in Fig.1. That part of the feature tree is just shown as below.

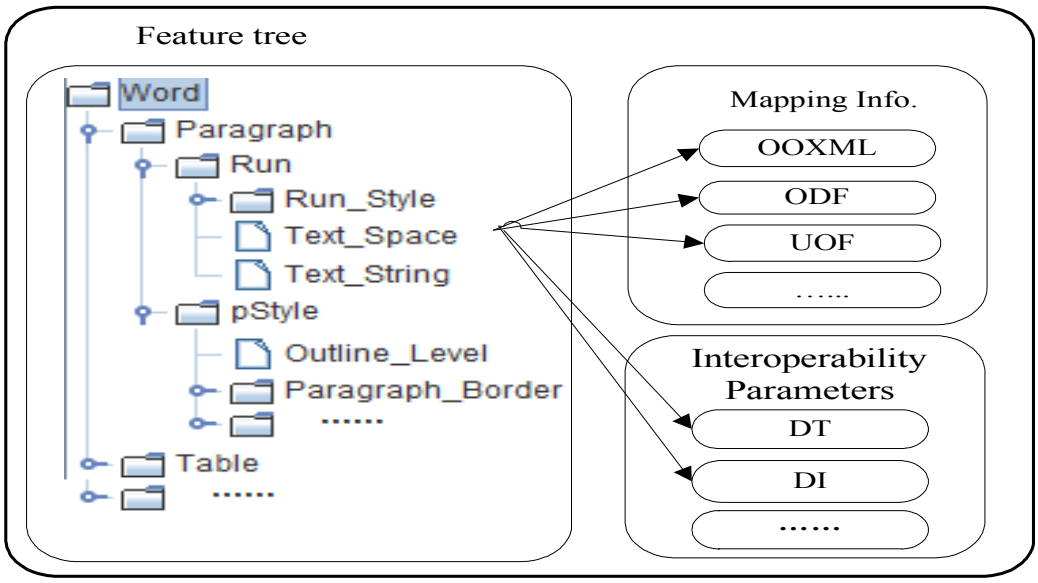

Figure 1: Structure of FDM

MI is used to specify how a feature is described in a certain document format. For example, MI in OOXML of feature *Text_Space* is its Xpath in OOXML documents like "/w:document/w:body/w:p/w:r/w:t/@xml:space="preserve"; therefore, MI can be used not only to detect whether a feature exists in an instance document but also detect how many times a feature occurs in the instance document.

Values of features' DI are often evaluated by experts in the field of document standards or office software. Generally, DI is divided into three degrees , specifically, the basic level, core level and deluxe level. The explanation of these levels is shown in Table 1. To get quantitative results, the three levels are designated by numbers of 3, 2 and 1 respectively. DI values in FDM are from experts.

Level	***Explanation***	***Value***
basic level	The essential features in office software	3
core level	Commonly used features in office software	2
deluxe level	Complex and not commonly used features in office software	1

Table 1: DI Degrees

The software vendors will ensure the features with higher rank to be realized in their office product at first. It is important for their products to pass the standard compatibility tests; therefore, the DI evaluation by experts is instructive.

Suppose that for a feature f_i, its DI from experts is $\boldsymbol{p(f_i)}$**.** There are two problems in the processing of experts' evaluation.

(1) The workload is huge. The amount of features is too large so that the artificial workload is too big; as a result, the experts may have different evaluation results for the same feature. It has to collect many experts' opinions and synthesize a final result.

(2) As the Results are subjective, they cannot be fit to any situation. For example, there are two features f_1 and f_2 which belong to the basic level and deluxe level respectively. Suppose the times of f_1 and f_2 occur in an instance document d is num_1 and num_2 respectively. Though $p(f_1)>p(f_2)$, there is $num_1<<num_2$. The actual situation and experts' evaluation is contradictory. In this case, the experts' value is not appropriate to evaluate these features' DI in the instance document.

In order to avoid the above problems, statistical methods are proposed and a system is given to evaluate the values of DI automatically.

3. Statistical Methods to Evaluate DI

Here, we use the usage frequency to evaluate DI. Bigger the usage frequency is, higher the DI value will be. The values of DI are computed from a set of sample documents by statistical methods objectively to avoid subjectivity from persons. To fit different requirements, three statistical methods are proposed in this section.

Without loss of the generality, suppose

(1) The set of sample documents is D, the number of elements in D is N, namely $|D|=N$. An arbitrary sample document $d_j \in D$, $1 \leqslant j \leqslant N$.

(2) The set of features extracted from D is F, the number of elements in F is $|F|=M$. An arbitrary feature $f_i \in F$, $1 \leqslant i \leqslant M$.

3.1 General Evaluation Method

The DI of f_i can be evaluated by Formula (3.1), where $I_j(f_i)$ is an agitation function, namely if f_i is used in d_j, then $I_j(f_i)=1$, otherwise $I_j(f_i)=0$.

$$p_1(f_i)=\frac{\sum_{j=1}^{N} I_j(f_i)}{N} \tag{3.1}$$

Formula (3.1) evaluates DI according to whether features are used in sample documents and its value is in the range of [0,1]. This method is simple to understand and use for the users; thus it is called the *General Evaluation Method*.

3.2. Enhanced General Evaluation Method

In order to evaluate DI more accurately by usage frequency of the features, Formula (3.2) is used to improve Formula (3.1).

$$p_2(f_i)=\frac{\sum_{j=1}^{N} N_j(f_i)}{\sum_{a=1}^{M}\sum_{j=1}^{N} N_j(f_a)} \tag{3.2}$$

where $N_j(f_i)$ is the times of feature f_i occurs in the sample document d_j.

Formula (3.2) not only examines whether a feature appears in the sample documents, but also counts its occurring times in all the sample documents; therefore, this method is called the Enhanced General Evaluation Method.

3.3 Specified Evaluation Method

DI from Formula (3.1) and (3.2) reflects the overall usage situation of features; however, sometimes, the overall results and the usage situation in a specific document may be biased.

For example, as to two features of $f_p, f_q \in F$, there is $p_1(f_p)>p_1(f_q)$ or $p_2(f_p)>p_2(f_q)$. But for a certain document d_j, f_p is not used, while f_q is used many times. Namely $N_j(f_p)<N_j(f_q)$. Obviously, the results from both of Formula (3.1) and (3.2) are not reasonable for d_j.

Formula (3.3) is given to evaluate DI of a feature f_i in a certain document d_j. The result reflects the usage frequency of f_i in d_j, and it is valid only for d_j; thus it is called *Specified Evaluation Method.* To be more simplified, $p_3(f_i)$ stands for $p_3(f_i, d_j)$ in the latter part of the paper.

$$p_3(f_i, d_j) = \frac{N_j(f_i)}{\sum_{a=1}^{M} N_j(f_a)} \tag{3.3}$$

4. Experiments

In order to analyze the three methods proposed in this paper and contrast them to experts' results, we designed and used Java to implement a prototype system, Statistical Tool of Usage Frequency of Features (STUFF). The design of STUFF and some experiment data are given in this section.

4.1 Design of A Prototype System

STUFF can automatically extract the set of features used in all the documents in *D* and compute DI values of the features based on statistical Formula (3.1)-(3.3).

The structure of STUFF is shown in Fig.2. A set *D* of sample documents in accordance with a specific document format (OOXML, ODF or UOF) are deemed as the input of the system. On the basis of MI in FDM, the Detector in STUFF extracts automatically the set of features used in each document instance and sums them up to the Feature Set *F*. At the same time, the Calculator counts the occurrence times of all the features and evaluates the values of DI according to formula (3.1)–(3.3).

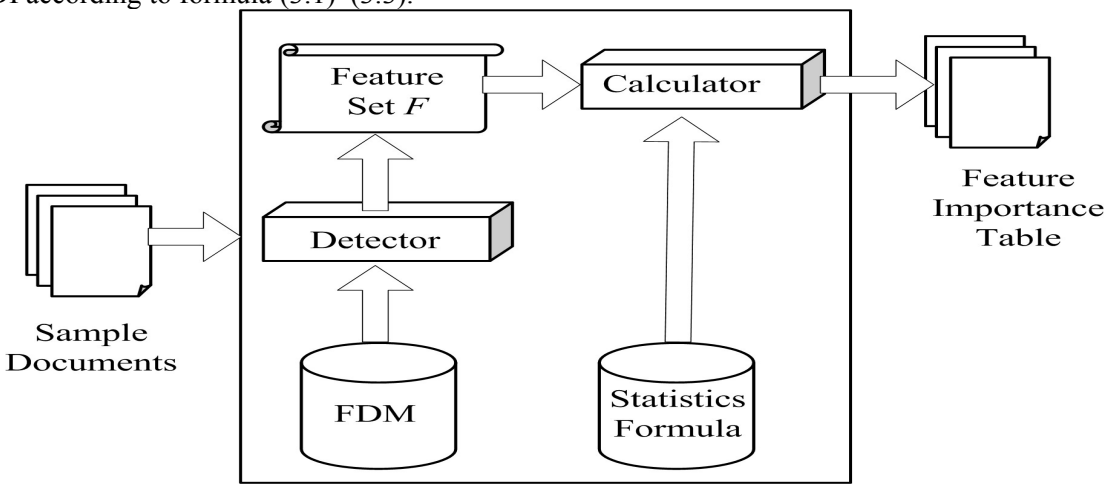

Figure 2: The structure of STUFF

4.2 Data Analysis

The following items are analyzed in this paper.

(1) According to the definitions of Formula (3.1), (3.2) and (3.3), their results are all in the range of [0,1], but the value of $p(f_i)$ is discrete as 1, 2 or 3. In order to achieve the consistency, $p_1(f_i)$, $p_2(f_i)$, and $p_3(f_i)$ should be mapping to degrees of 1, 2 and 3.

(2) From certain perspective, all of $p(f_i)$, $p_1(f_i)$ and $p_2(f_i)$ are the common degree of DI from the whole level. It should be validated whether the three values are consistent or not.

(3) How to make a choice if $p(f_i)$, $p_1(f_i)$, $p_2(f_i)$, and $p_3(f_i)$ are not equal.

As our work is only based on the document formats, the content and language of documents do not influence the results. We collected 500 word-processing documents randomly to be the set $D(|D|=N=500)$ of sample documents. The set *D* includes journal papers, dissertations, letters, and office papers. The Feature set *F* extracted from the samples includes 311 features ($|F|=M=311$). And then the experimental results are obtained automatically based on *D* and *F* by STUFF. Based on the experiment data, the three items mentioned above are solved.

(1) Degrees of statistical results

Suppose $count_1$, $count_2$ and $count_3$ are the numbers of features whose $p(f_i)$=3, $p(f_i)$=2 and $p(f_i)$=1 respectively. After counting based on D and F, there is $count_1$: $count_2$: $count_3$ =39:3:1.

The range of [0,1] is divided into three sub-regions and each sub-region corresponds to a degree of 1, 2 or 3. The boundaries of sub regions are decided by the distribution of $p_k(f_i)$ in [0,1] for k={1,2,3}, according to the ratio of 39:3:1.

Suppose the discrete degree of $p_k(f_i)$ is $dp_k(f_i)$, and the boundaries are r_1 and r_2. The steps to grade $p_k(f_i)$ to $dp_k(f_i)$ are as follows:

(a) Compute the values of $p_k(f_i)$ for $\forall f_i \in F$, $1 \le i \le M$.

(b) Sort all the vales of $p_k(f_i)$ from small to big, and get a sequence A as

$$A=\{a_i \mid a_i<a_j \text{ , if } i<j, 1 \le i,j \le M\} \tag{4.1}$$

(c) Compute the boundaries

For (m=1;$m \le M$; m++)

{

if (m/M = =1/(39+3+1))

$r_1 = a_m$;

else (m/M = =(1+3)/(39+3+1))

$r_2 = a_m$;

}

(d) Let the sub regions are [0, r_1], [r_1, r_2] and (r_2,1].

(e) If $p_k(f_i) \in [0, r_1]$, then it shows the usage probability of f_i is very low; therefore, specify $dp_k(f_i)$=1. We can use the rules as follows.

For each f_i

$$dp_k(f_i)=\begin{cases} 1, if\ 0 \le p_k(f_i) < r_1 \\ 2, if\ r_1 \le p_k(f_i) \le r_2 \\ 3, if\ r_2 < p_k(f_i) \le 1 \end{cases} \tag{4.2}$$

The sub-regions for $p_k(f_i)$, k=1,2,3 and their corresponding degrees are given in Table 2. For example

- if $p_1(f_i)$=0.003, then it is graded as $dp_1(f_i)$=1.
- if $p_2(f_i)$=0.003, then it is graded as $dp_1(f_i)$=3.
- if $p_3(f_i)$=0.003, then it is graded as $dp_1(f_i)$=2.

dp_i	**Degrees**		
p_i	**1**	**2**	**3**
p_1	[0,0.1)	[0.1,0.5]	(0.5,1]
p_2	[0,0.0001)	[0.0001, 0.002]	(0.002,1]
p_3	[0,0.0001)	[0.0001, 0.005]	(0.005,1]

Table 2: Mapping Rules of Degree

(2) Consistency Analysis

Statistical values $p_k(f_i)$, k=1,2,3 are graded to be $dp_k(f_i)$. So $dp_k(f_i)$ and $p(f_i)$ can be compared with each other easily. Some results from samples are given in Table 3.

Features	Leaf Features	The semantic of features	*p*	dp_1	dp_2	dp_3
Meta	Author	the document producer that was used to create or last modify the document	2	1	1	1
	Company	the company which the document is belong to	3	1	1	1
	Edit_Times	the number of times a document has been edited	3	3	2	3
Run_Style	Text_Bold	Text within the feature should be bold.	3	3	3	3
	Run_Border_Style	Describe the border style of a run.	1	1	1	1
	Font_family	Specify the font family.	3	3	3	3
	Font_size	Specify the font size.	3	3	3	3
Paragraph_Style	Line_Space_AtLeast	Specify the line space at least.	3	3	3	2
	Line_Space_Auto	Specify the line space automatically.	3	3	3	3
	First_Line_Indent_Absolute	Specify the first line indent of a paragraph with an absolute distance.	3	3	3	3
	Paragraph_Before_Auto	Specify the space before a paragraph automatically.	3	2	2	2
Layout	Left_Margin	Specify the left margin.	3	3	2	3
	Outline_Level	Specify the outline level.	3	2	2	2
	DocGrid_Line	Specify the number of document columns.	3	3	2	2
Table	Table_Fill_Color	Specify the color fills a table	3	2	2	2
	Table_Border_Top_Style	Specify the style of a talbe's top border.	3	3	2	3

Table 3: Degree Values of DI

For $\forall f_i \in F$, if $dp_1(f_i)=dp_2(f_i)=p(f_i)$, then let $a_i=1$ and $b_i=0$; otherwise let $a_i=0$ and $b_i=1$. Based on the experimental results, the following results are gotten.

$$A=\frac{\sum_{i=1}^{M} a_i}{M}\times 100=\frac{134}{311}\times 100=43.09\ \ \% \tag{4.3}$$

$$B=\frac{\sum_{i=1}^{M} b_i}{M}\times 100=\frac{177}{311}\times 100=56.91\ \ \% \tag{4.4}$$

From the experimental results, there is difference between values from the experts and the statistical methods. For example, the feature of metadata "*Company*", the experts believe it very basic and important for document formats and give its DI to be 3, while it is seldom used in sample documents and all the statistical values are 1.

The reason is that experts evaluate DI not only according to the usage frequency of features but also other factors such as the definition of document format standards and the standards compatibility testing, etc. However, the statistical methods compute DI only according to the usage frequency of features. Though the values of DI are affected by the samples, the influence should be small if the sample set is large enough. Therefore, the difference between experts and statistical methods does not explain which kind of evaluation method is not reasonable, while embodies the objective situation.

(3) Rules based on voting

Facing $p(f_i)$, $dp_1(f_i)$, $dp_2(f_i)$ and $dp_3(f_i)$, users should make a choice according to their requirements.

Distinguishingly, $dp_3(f_i)$ is obtained from a specific document and the results are not universal for other documents. Even the values of $dp_3(f_i)$ obtained from two instance documents are very different; therefore, when the need is to highlight the impact of features in a specific document, $dp_3(f_i)$ is the most suitable.

While all of the $p(f_i)$, $dp_1(f_i)$ and $dp_2(f_i)$ reflect the overall effect, how to make a choice? The final result should be one of the three kinds of values or be an integrated value based on the several methods. Here we give rules based on voting.

As to an arbitrary $f_i \in F$ and its value set $V(f_i)=\{dp_1(f_i), dp_2(f_i), p(f_i)\}$, where $V(f_i)$ is a unordered set. The final value $v(f_i)$ is decided by the rules in Table 4. Namely, $v(f_i)$ is the value which occurs more times in $V(f_i)$. If there are three different values, namely $V(f_i)=\{1,2,3\}$, then let $v(f_i)=2$, the median value.

Value 1	Value 2	Value 3	Final value
1	1	2 or 3	1
2	2	1 or 3	2
3	3	1 or 2	3
1	2	3	2

Table 4: Rules Base on Voting

5. Conclusion

Degrees of importance (DI) of document features are very useful in the field of document processing. The workload of evaluating DI is too large by human, and the results from experts cannot be fit to any application requirement.

In this paper, three kinds of statistical methods are proposed to evaluate the values of DI. After counting the usage frequency of features in a large set of sample documents, we can obtain results to evaluate the DI of features objectively. It is helpful to avoid the subjective factors in the processing of expert evaluation. Additionally, based on FDM, STUFF is implemented to do the statistics work automatically. Compared with the prior manual evaluation, it can greatly improve the work efficiency.

References

[1] *Information technology-Open Document Format for Office Applications (OpenDocument)* v1.0[S]. ISO/IEC Std. 26300. 2006.

[2] *Information technology - Office Open XML file formats*[S]. ISO/IEC Std. 29500. 2008.

[3] *Specification for the Chinese office file format*[S]. GB/T Std. 20916. 2007.

[4] X. Hou, N. Li, H.B.Yang, Q. Liang. C*omparison of Wordprocessing Document Format in OOXML and ODF*[A]. Proc. of Sixth International Conference on Semantics, Knowledge and Grids, 2010.

[5] S.Hiser. *Achieving Openness: A Closer Look at ODF and OOXML*[EB/OL]. Avaliable at http://onlamp.com/pub/a/onlamp/2007/06/14/achieving-openness-a-closer-look-at-odf-and-ooxml.html.

[6] K. P. Eckert, N. Li, X. Hou, D. S. NAM. ISO/IEC 29166:2011 *Information technology--Document description and processing languages--Guidelines for translation between ISO/IEC 26300 and ISO/IEC 29500 document formats*[TR]. ISO. 2011.

[7] N. Li, Q. Liang, X. Hou. *Interoperability Measurement of Documents*[J]. Chinese Journal of Electronics, 21(1):37-41(2012).

[8] K. P. Eckert, G. Kerstin. *Feature based document profiling - a key for document interoperability?* [TR]. 2012. availible at http://www.interoperability-center.com/c/document_library/get_file?uuid=8c99ebc0-c33b-4e45-bdd2-7456e2277411&groupId=12725

[9] Y.W. Gao, X. Hou and N. Li. *An implementation of documents interoperability measuring system*[J]. Applied Mechanics and Materials. vol.385-386 :1764-1770(2013).

[10] X. Hou, N. Li and Q. Liang. *Ontology based measurement model for document interoperability*[J]. Computer Engineering and Design. 35(10):3467-3471(2014) (In Chinese)