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.
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.
Statistical Methods to Evaluate Important Degrees of Document Features

Xia Hou

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.

<table>
<thead>
<tr>
<th>Level</th>
<th>Explanation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic level</td>
<td>The essential features in office software</td>
<td>3</td>
</tr>
<tr>
<td>core level</td>
<td>Commonly used features in office software</td>
<td>2</td>
</tr>
<tr>
<td>deluxe level</td>
<td>Complex and not commonly used features in office software</td>
<td>1</td>
</tr>
</tbody>
</table>

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 \( p(f_i) \). 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 \( \text{num}_1 \) and \( \text{num}_2 \) respectively. Though \( p(f_1) > p(f_2) \), there is \( \text{num}_1 << \text{num}_2 \). 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 $d_j \in D$, $1 \leq j \leq N$.

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

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 $I_j(f_i)=0$.

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

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).

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

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 $f_r, f_q \in F$, there is $p_1(f_r)>p_1(f_q)$ or $p_2(f_r)>p_2(f_q)$. But for a certain document $d_j$, $f_r$ is not used, while $f_q$ is used many times. Namely $N_j(f_r)<N_j(f_q)$. 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.
\[ p_3(f_i, d_j) = \frac{N_j(f_i)}{\sum_{a=1}^{M} N_j(f_a)} \]  

(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).

![Diagram of STUFF](image)

**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 \( p(f) \) is discrete as 1, 2 or 3. In order to achieve the consistency, \( p_1(f), p_2(f), \) and \( p_3(f) \) should be mapping to degrees of 1, 2 and 3.

(2) From certain perspective, all of \( p(f), p_1(f), \) and \( p_2(f) \) 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 \( p(f), p_1(f), p_2(f), \) and \( p_3(f) \) 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 \( \text{count}_1, \text{count}_2 \) and \( \text{count}_3 \) 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 \( \text{count}_1; \text{count}_2; \text{count}_3=39:3:1 \).

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 \( \forall f_i \in F, 1 \leq i \leq M \).

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

\[
A = \{a_i | a_i < a_j, if i < j, 1 \leq i,j \leq M \} \quad (4.1)
\]

(c) Compute the boundaries

For \( m=1; m \leq M; m++ \)

\[
\begin{align*}
& r_1 = a_m; \\
& r_2 = a_m; \\
& \text{if ( } m/M = 1/(39+3+1) ) \\
& \text{else ( } m/M = (1+3)/(39+3+1) ) \\
& \end{align*}
\]

(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 \( f_i \)

\[
\begin{align*}
& 1, if 0 \leq p_k(f_i) \leq r_1 \\
& 2, if r_1 \leq p_k(f_i) \leq r_2 \\
& 3, if r_2 < p_k(f_i) \leq 1 \\
& \end{align*}
\] 

(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 \( p_1(f)=0.003 \), then it is graded as \( dp_1(f)=1 \).
- if \( p_2(f)=0.003 \), then it is graded as \( dp_2(f)=3 \).
- if \( p_3(f)=0.003 \), then it is graded as \( dp_3(f)=2 \).

<table>
<thead>
<tr>
<th>( p_i )</th>
<th>( dp_i )</th>
<th>Degrees</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_1 )</td>
<td>[0.0, 0.1)</td>
<td>[0, 0.1, 0.5]</td>
</tr>
<tr>
<td>( p_2 )</td>
<td>[0.000001, 0.0001)</td>
<td>[0.0001, 0.002]</td>
</tr>
<tr>
<td>( p_3 )</td>
<td>[0.000001, 0.0001)</td>
<td>[0.0001, 0.005]</td>
</tr>
</tbody>
</table>

Table 2: Mapping Rules of Degree

(2) Consistency Analysis

Statistical values \( p_k(f), k=1,2,3 \) are graded to be \( dp_k(f) \). So \( dp_k(f) \) and \( p(f) \) can be compared with each other easily. Some results from samples are given in Table 3.
Statistical Methods to Evaluate Important Degrees of Document Features
Xia Hou

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

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.

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

$$B = \frac{\sum_{i=1}^{M} b_i}{M} \times 100 = \frac{177}{311} \times 100 = 56.91\%$$  \hspace{1cm} (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

\[ \text{Page 7} \]
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 an 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.

<table>
<thead>
<tr>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Final value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2 or 3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1 or 3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1 or 2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

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


