Re-flowable Document Structure Understanding by Comprehensive Use of Features and Rules

Hai-li HAO*, Ning LI, Ying-ai TIAN and Si GENG

Computer School, Beijing Information Science and Technology University, Beijing 100101, China

*Corresponding author

Keywords: Document understanding, Document component identification, Re-flowable document.

Abstract. Aimed to improve the shortcomings in the previous re-flowable document structure understanding that component features are not fully utilized in component identification, this paper proposed a new method to understand documents in combination with features and grammatical rules. In the method, two vectors are used, the first one is the format vector representing the format features, such as fonts etc., the second one is the content vector representing text features such as keywords etc. Then the components to be identified are compared with the candidates by measuring the distance between the vectors with different weights. The experiment results show that this method can effectively improve the accuracy of component identification, and in turn improve the accuracy of whole document understanding.

Introduction

With the development of social information, various documents will be produced in daily work and life. Streaming documents cover a large amount of information. Compared with ordinary documents, re-flowable documents have specific format information appended to the text content. Document components refer to substructures with relatively independent semantics and independent functions in the document format. Different document components have different format information, and there are hierarchical nesting relationships between document components. The extraction of re-flowable document format information helps the computer to better understand the document and automatically typeset the re-flowable document, so it is valuable. Most of the past re-flowable document studies used a template-based approach[123], the identification of document titles generally adopts the regular matching method [4], but the limitation is large, for example, the template is difficult to describe the complex document structure.

Document understanding includes layout document understanding, re-flowable document understanding, etc. Layout document understanding has accumulated a lot of research results [5678], the layout understanding is mainly aimed at fixed format documents, divided automatically and judged the regional and positional relations of elements in the layout, and then mapping the physical structure of the layout to logical structure. Literature [5] proposed the SVM learning model algorithm to get the correct layout results, and solved the problem that the optimal parameters were not accurately obtained in the traditional layout analysis algorithm. There are also some typographical understandings of specific format content, such as tables and graphs [9]. Web pages contain a large amount of semi-structured information [10], and Web document understanding is mainly applied to Web information extraction. Literature [11] used heuristic method to deduce the hierarchical structure on the page according to the font size and indent distance of each part.

This paper focuses on the identification of document components in re-flowable document understanding [12][13]. Document component contains the content features of a text, feature selection affects the result of component recognition and the accuracy of the structure recognition. However, the features commonly used in component identification include content feature, format feature, object feature, etc. Such as KanYun Ji [14] tested the keywords like "abstract", "catalog", and the serial number "x.x", "X.X.X" to determine the logic tag document components, the method based
on the content features is simple and rapid, but often rely on document type resulting in poor
generality. The method based on VSM [15] made full use of the format features, but there are many
features in format, including qualitative components (such as font) and quantitative (such as name), it
is difficult to find the proper transform method and weight when non-dimensionalized. Pengxin [16]
solved the problems by combining the n-gram index and the membership recognition document
component of pattern recognition in the spelling correction. However, this method fails to use the
content features that contribute to structural identification n. This paper analysis the format of the
component from features and content, we calculate the component score, two kinds of characteristics
calculated by weighted final component recognition as a result, the final document using document
typesetting rules is logical structure. The component scores of the two characteristics are calculated
respectively, and the final component identification results are obtained by weighted calculation.

The Construction of the Document Layout Rules

Document layout rules describe the document structure based on document components, the types of
document artifacts can be divided according to the format function points, the logical structure of
specific documents and the classification of rhetorical structures. In this paper, we use the idea of
Document Component Ontology (DoCO) [17], and divide the document components according to the
function expressed in paragraphs. Document structure is usually a tree structure, the construction of
the document rules through hierarchical relations describe the structure of the document, including
the order that component appear, the number of occurrences, nested structure and so on. XML
Schema is a way of defining a vocabulary, making it easy to represent tree information structures and
clearly describing the layout rules of a document. Therefore, this paper uses the layout rules based on
XML Schema to describe the rules that different components compose the document. Take the more
common degree thesis in the document as an example.

![Diagram](image)

Figure 1. The rules of document typesetting based on XML Schema.

In addition to representing the document structure in the typesetting rules, the layout format of each
component is also described, which is divided into the format features and content features. The
format features include font, size, character pattern, alignment, text-indent, spacing and line spacing
between segments. The content feature contains keywords or specific formats for the component.

Document Component Identification

In the process of component identification, firstly, the format vector is established to represent the
character of components such as font. Extracting key words and other content features in document
components as key words, the two features of the identified components and the scores of the
candidate components are calculated respectively.

Document Component and Feature Analysis

In streaming documents, the basic layout elements are sentences and paragraphs. In addition, charts
and other objects are included. Therefore, the document artifacts can be divided into two categories:
1) basic components refer to text class members, such as summaries and titles, which are text
paragraphs as basic units; 2) special components, refer to non-verbal content or components
composed of complex content, such as pictures and forms. The document component contains many features, and the common features are shown in section 3.2. This paper analyzes the three aspects: format feature, content feature and object feature.

**Format Feature.** Different document components are formatted differently, which helps to distinguish between different document components. For example, paragraphs with bold font and font size are usually paragraph headings. This article selects the features that are often used to set up the components for analysis, including font, fontSize, fontShape, alignment, firstLineIndent, lineSpacing, spaceBefore, spaceAfter. Therefore, according to the format information of the paragraph, the corresponding components of the paragraph can be preliminarily determined.

**Content Features.** By analyzing the content of the document, it is found that the document has its fixed document format, which contains specific keywords, and has a specific section number format, etc.

Keyword information: For example, the document component of the graduation thesis contains the "key words", "directory", "acknowledgement", "appendix" and other key words, these document components which are usually separated into sections only contain these keywords, other words are not including, and are separated into sections. In addition, in the following picture appeared in the "Fig" followed by the Numbers in the form of the paragraph beginning is commonly figure, appear above the Table with a "Table" followed by the Numbers in the form of the paragraph beginning is generally for the Table.

Specific section number format: Mainly used to judge the title level of the paper. For example, the number format of the first level title is "Chapter n XXX" or "n XXX", and the number format of the secondary title is "N.N XXX". Or Arabic numerals (1,2,3...) N can only be represented by Arabic numerals.

**Feature-based Component Identification**

This paper uses the format features and content features to identify the components. In terms of format features, this paper extends the types of identifying components on the basis of document component identification based on the format index [16]. According to the current document format information of components, the use of the eight format characteristics of high degree of differerntiation, including: font name, fontSize, fontShape, alignment, firstLineIndent, lineSpacing, spaceBefore, spaceAfter. According to the format information to establish a standard template document and document under test dictionary, dictionary recorded in written documents and document templates each format features appear in their respective document values, each value corresponding to the different components, the component with the highest score is the component corresponding to the paragraph by calculating the scores on each component. The concrete algorithm based on the format feature identification component is shown below.

<table>
<thead>
<tr>
<th>Algorithm: Component Recognize</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Let c be the component to be identified</td>
</tr>
<tr>
<td>2. Let C be the component set</td>
</tr>
<tr>
<td>3. Let f be a feature</td>
</tr>
<tr>
<td>4. Let F be the feature set</td>
</tr>
<tr>
<td>5. Let dw be the words of document dic that will be identified</td>
</tr>
<tr>
<td>6. Let w be the words of standard document</td>
</tr>
<tr>
<td>7. for each c ∈ C</td>
</tr>
<tr>
<td>8. for each f ∈ F</td>
</tr>
<tr>
<td>9. for each dw ∈ dic_c</td>
</tr>
<tr>
<td>10. if V(c) = dw and s(dw,w)=max{ (s(dw, the branches of dw)) }</td>
</tr>
<tr>
<td>11. score(c)= s(dw,w)* e(c, the index of dw)* w(c,f)</td>
</tr>
<tr>
<td>12. End if</td>
</tr>
<tr>
<td>13. score(c)=∑ score(c)</td>
</tr>
<tr>
<td>14. if score(c)=Max(score(c))</td>
</tr>
<tr>
<td>15. return c</td>
</tr>
<tr>
<td>16. end for</td>
</tr>
</tbody>
</table>
Among them, \( V_f(c) \) is the value taken by component \( c \) on a certain feature, \( s(dw,w) \) is the membership degree of the word \( w \) in the lexicon of the document dictionary, and \( e(c, \text{the index of } dw) \) is the occurrence number of component \( c \) in the word \( dw \) of the document to be tested. \( W(c,f) \) is the weight of component \( c \) on feature \( f \), and \( \text{score}(c) \) is the score of component \( c \) on some feature \( f \), \( \text{score}(c) \) is the total score of component \( c \) in all format features.

In terms of content features, different document components have different content features. This paper uses a variety of content features to identify document components. Five features are used in this paper. Which is: object feature, keyword, chapter number format, punctuation and sentence number limitation. The content feature represented as a vector, and then the vector is different or manipulated with the vector of all the components in the standard template, namely the normalization of the vector.

Finally, the score of the identified paragraph in each component is calculated, and the highest score is selected as the corresponding component of the paragraph. The calculation formula is as follows.

\[
\text{Result} = a \cdot w_1 + b \cdot w_2 + c \cdot w_3 + d \cdot w_4 + e \cdot w_5
\]

(1)

In equation (1): \( a, b, c, d, e \) respectively represent the object features component, the keyword component, the component of the chapter number format, the punctuation component, and the sentence number limiting component; the following letters \( w_1 - w_5 \) represents the weight of all the signatures. For the weight setting, this paper uses the idea of greedy algorithm to decompose the whole problem into multiple stages, and through a series of local optimal solutions to achieve the overall optimal solution. According to the contribution of each content feature to document component recognition, the weight of each feature is adjusted in turn, so that the recognition result is optimal, and the weight is 0.4, 0.25, 0.25, 0.05 and 0.05.

Based on the format feature to identify the document components, when two different document components are formatted in the same format, then two candidate components will be identified, which will affect the subsequent document structure identification results. Based on the content feature, the document component can be identified, which is applicable to documents with a specific structure. Once the content features of the document are not present, the identification results will be affected, and even the unrecognized cases will appear. So, in this paper, we consider the format and the features, the weighted calculation of its features, and the point of the final result of the final point of the score. Here's the formula:

\[
\text{FinalResult} = \text{formatScore} \cdot w_6 + \text{contentScore} \cdot w_7
\]

(2)

In equation (2), \( \text{formatScore} \) represents the score based on the format feature recognition, \( \text{contentScore} \) is the score based on the content feature, the following letter \( w_6 \) and \( w_7 \) respectively represent its weight, and the weight is determined by the experiment result, which is 0.5 and 0.5.

Experimental Results and Analysis

<table>
<thead>
<tr>
<th>Document component</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>98.56%</td>
<td>98.16%</td>
<td>98.36%</td>
</tr>
<tr>
<td>Body</td>
<td>97.50%</td>
<td>99.56%</td>
<td>98.52%</td>
</tr>
<tr>
<td>Keyword</td>
<td>99.56%</td>
<td>99.56%</td>
<td>99.56%</td>
</tr>
<tr>
<td>Heading 1</td>
<td>98.06%</td>
<td>99.56%</td>
<td>98.80%</td>
</tr>
<tr>
<td>Heading 2</td>
<td>98.33%</td>
<td>98.46%</td>
<td>98.39%</td>
</tr>
<tr>
<td>Heading 3</td>
<td>99.27%</td>
<td>99.27%</td>
<td>99.27%</td>
</tr>
<tr>
<td>Picture</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Picture caption</td>
<td>97.89%</td>
<td>98.27%</td>
<td>98.08%</td>
</tr>
</tbody>
</table>

294
Based on the feature recognition document component stage, this paper adopts the degree thesis of Beijing Information Science and Technology University as the experimental data set. The experimental data set is now available on GitHub (https://github.com/COSLab) and gradually expands the data volume. There are more than ten different documents that are identified in the experiment, see table 1. This experiment adopts the accuracy rate and recall rate and F values to review document component recognition results. The identification results are shown in table 1.

As can be seen from table 1, the accuracy and recall rate of the document component algorithm combined with multiple feature recognition are relatively high, indicating that the method can accurately identify the corresponding components of the paragraph and find the candidate paragraphs of a document component as much as possible. Among them, the accuracy of the graph and the table is low, because the graph and the table are based on the keyword "graph" or "table" to carry out the content feature identification, but because the document writer often writes the number without writing the key words in the drawing or table problem, the paragraph that originally belongs to the graph or the table problem is misjudged as the text.

The comparison results of document component recognition algorithm based on multiple features and the document component recognition algorithm based on VSM [15] and the document component identification method based on the format index [16] are tested in the same experimental set as shown in figure 2.

As can be seen from figure 2, the accuracy of component recognition algorithm based on VSM is the lowest, because the algorithm can only be compared with the preset logic label, and can’t be compared with any logical label, resulting in poor effect. The accuracy of the component recognition algorithm based on the format index is higher than that of the VSM based component recognition algorithm. The reason is that the algorithm can compare the paragraphs with any logical label without having to set the logical label in advance. The recognition algorithm, which combines multiple characteristics, has improved the accuracy of the document building recognition. Besides, it relates to two other methods.

### The Conclusion

Based on the analysis of the method of document understanding at home and abroad, this paper proposes a new method based on fusion feature and grammar rules for the lack of feature analysis in the process of document component identification. According to the various information contained in the document component, the fusion format feature and content feature are used as document component identification features to improve the accuracy of the component recognition and lays the
foundation for the later document format optimization. The experimental results show that the method is more accurate than other methods based on feature recognition. However, the weights of various features are statistically analyzed by using multiple features to identify the document structure. In order to identify the document components more accurately, the required features can be found automatically by machine learning method, which is the direction of future document component recognition. Meanwhile, it is also important to extend the identification document types in the future.

Acknowledgement
This research was supported by the Intelligent Analysis and Optimization Method for Re-flowable Documents through the National Natural Science Foundation of China.

References


