DOI: 10.23977/acss.2022.060711
ISSN 2371-8838 Vol. 6 Num. 7

Adaptive Selection Algorithm of Paragraph Selection Structure Based on Complex Data Type

Aoxue Han[1,a,*], Zhinan Lin[2,b]

[1]Hunan University of Science and Technology, Xiangtan, China
[2]Hunan Vocational Institute of Technology, Xiangtan, China
[a]binbin135197@163.com, [b]409435402@qq.com
**Corresponding author*

Keywords: Document Generation, Complex Data Type, Selection Structure, Adaptability

Abstract: In the automatic generation of standard documents, the replacement method of template is cumbersome and difficult to modify and maintain in the later stage when dealing with the selection structure in the paragraph template. To solve this problem, an adaptive paragraph selection structure selection algorithm based on complex data types is proposed. On the basis of the tree structure, each node is abstracted into a class, selection nodes are set, and attributes and methods are set for the selection nodes. The deletion and retention of nodes in the paragraph template are completed through the attributes and methods of the selection nodes. The experiment shows that it has a good effect in paragraphs with multi branch sentences.

1. Introduction

Normative documents are professional documents with strong standardization, which can be automatically generated by computer, such as medical examination report [1], weather detection report [2], etc. For automatic generation of standard documents, templates are usually used to control the content and structure of documents. The document template consists of fixed content and dynamic content. The fixed content is unchanged, while the dynamic content is determined by the data provided. When the data changes, the dynamic content will also change.

The structure of normative documents is divided into outline, chapter and paragraph. The corresponding document templates are outline template, chapter template and paragraph template. Paragraph is the most basic unit of a document, and the paragraph template is the basis of the document template. A document consists of multiple paragraphs, which may include text, formulas, pictures, and tables. There may be certain relationships between paragraphs and within paragraphs, such as selection relationship, causal relationship, parallel relationship, etc., which can be realized by setting different template structures.

Selection structure is an important part of document generation. The paragraph is composed of multiple sentences. When a paragraph is composed of sentences, you need to select the sentence and select the desired sentence from all the options. Sentences contain phrases, which also need to be selected when automatically generating sentences. The selection of paragraphs to sentences and the selection of sentences to phrases belong to the selection structure, which is realized through the control template in the automatic document generation mode.

For the generation of paragraph content, the replacement method is usually used. At present, the replacement method of Word document template is commonly used[3-8]. This method takes Word document as template directly, saves the pre prepared Word document with label as XML file, and then modifies the dynamic content of XML file through template engine to form a new template. Then parse the new template, replace the label and data, and re compress them into Word documents. The replacement method is easy to make templates, but the document structure is fixed and cannot be modified when making templates. When the content structure of the document changes greatly, the template can only be made again, which takes a lot of time. In the paragraphs with a large number of selected structures, each wrong selection will lead to changes in the content or structure of the template, and the template may need to be made again.

Aiming at the above problems, an adaptive selection algorithm of paragraph selection structure based on complex data type is proposed, which allows the dynamic content in the template to modify the template structure through the data in the database, and automatically forms a new template. Compared with the traditional Word template method, this method saves the steps of manually modifying the template structure and replacing labels when there is a selection structure in the paragraph, and can meet the needs of actual standardized documents, which speeds up the efficiency of automatic generation of paragraph documents.

2. The Implementation Method of Selecting Structure in Word Document Template Replacement Method

The automatic generation of Word documents is to parse Word documents into XML files and modify XML files as templates. Extensible Markup Language (XML for short) is a markup language and a way to store data in simple text format. The automatic generation of paragraphs is to use XML files as paragraph templates. The entire paragraph can be viewed as a node tree, in which elements, text, and attributes are considered nodes. In the OOXML (Office Open XML Specification) format, the body file document is extracted from the Word document In XML, it can be roughly divided into paragraph element node, statement element node and text element node, among which the text element node also contains child nodes: text node.

Paragraph selection structure is divided into sentence selection structure and phrase selection structure, which are paragraph to sentence selection and sentence to phrase selection respectively. The statement selection structure in XML format is the selection of statement element nodes. The statement selection in a paragraph is completed by deleting statement element nodes. The phrase selection structure is to select a part of the text in the text node, set the label, and then replace the label.

The statement selection structure is implemented through the control node, but XML is only a way to store data, and external engines are needed to add, delete, modify, and query nodes. The expression of statement selection structure in XML template is to select one or more statement nodes. By traversing all nodes, you can find the node to be selected according to requirements, and delete all other nodes and their children. The operation of this method is very simple, but it requires manual determination of options and deletion of unnecessary options, which is not flexible enough. If you make mistakes in modifying the template, you need to regenerate the XML template, which is very tedious.

The statement selection structure is implemented through the control node, but XML is only a way to store data, and external engines are needed to add, delete, modify, and query nodes. The expression of statement selection structure in XML template is to select one or more statement nodes. By traversing all nodes, you can find the node to be selected according to requirements, and delete all other nodes and their children. The operation of this method is very simple, but it requires manual determination of options and deletion of unnecessary options, which is not flexible enough.

If you make mistakes in modifying the template, you need to regenerate the XML template, which is very tedious.

```
Begin
docXML ← genDocumentXML(eFile,dataMap)
//Generate document.xml file
Paragraph ← getParagraph(docXML)
//Get the paragraphs in the document.xml file
Tags ← getTags(Paragraph)
//Get all tags in the paragraph
  for i from 0 to Tags.size
      if  dataMap.containsKey(Tags[i])
//Traverse all tags
          then docXML ← replaceDocumentXML(docXML,dataMap)
//Replace tags in the document.xml file
return docXML
End.
```

Word document template replacement method uses XML file as a template to realize paragraph selection structure. The biggest advantage is that the template is simple to make, and the original template can be made directly through Word. As a way of storing data, XML needs to be modified through an external engine, which is cumbersome to operate. This method has already determined the dynamic content when making the template, and the template structure is not flexible enough.

3. Adaptive selection method based on complex data types

In order to solve the problem of paragraph selection structure in the template replacement method of Word documents, this paper constructs an adaptive selection algorithm based on complex data types. Complex data types are also called reference data types. When storing, variables store only addresses, while simple data types store values in variables. The essence of a class is a complex data type, which encapsulates properties and methods. In this method, the node tree is directly used as the paragraph template, and each node is treated as a complex data type class. At the same time, a new node type is set to select the basic node. The new node is named as the selection node and set as the root node in the selection structure.

Adaptive selection algorithm is to change the structure of the node tree by selecting the attributes and methods carried by the node when traversing the template. The selection node contains the query string, the selection parameter attribute, and the selection node method. Instantiate the node selection and use the node selection method to complete the selection of options by querying the string and selecting parameters, and operating the transformation node tree, as shown in Figure 1.

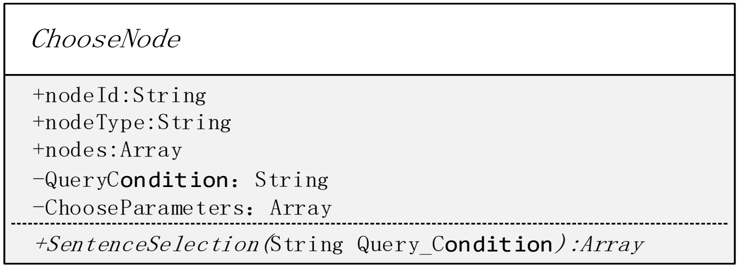


Figure 1: Select Node Class Diagram.

Query string attributes are spliced from table names, column names and query conditions in the database. After splitting it, use these three parameters to find all eligible selection parameters in the database. The selection parameter attribute is the ID of the sub node to be selected. It is an array, which can contain multiple values, one value or no value. It corresponds to three options: multiple selection, single selection and no selection. The selection structure method is to traverse the sub nodes of all the selected nodes, retain the sub nodes represented by all the selection parameters, and delete all the sub nodes not represented by the selection parameters.

The node tree of a paragraph template consists of paragraph nodes, sentence nodes and text nodes. The paragraph node is the root node, and the sentence node is the parent node of the text node. The sentence is composed of fixed text and dynamic text in one text node. The selection structure in a paragraph can be divided into sentence selection structure and phrase selection structure, both of which can be realized by selecting nodes. The core algorithm is as follows:

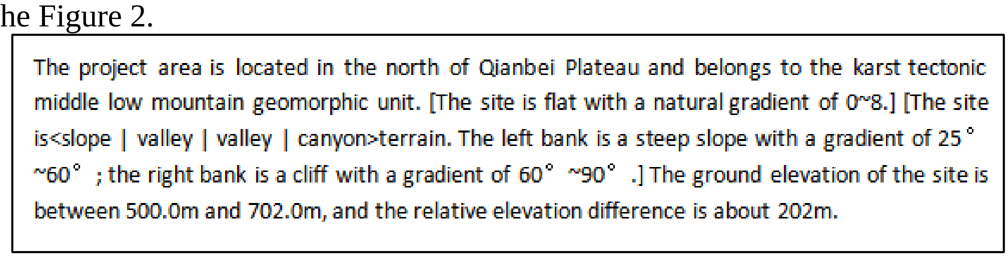
```
Begin
if  ChooseNode.nodeType=”Choose_node”
//The node type is selected node
        QueryCondition ← ChooseNode.QueryParameters.Split(',')
//Split the query string to get the table name, column name and query criteria
        ChooseParameters ← query(QueryCondition)
//Get the selection parameters according to the query criteria
        then m ← 0
        for i from 0 to ChooseNode.Nodes.length
//Traverse all child nodes of the selected node
        if !ChooseParameters.Exists(ChooseNode.Nodes[i].id)
  //Select whether the parameter has the ID of the current child node
                      then Choose_node.Nodes[i-m].Remove()
                      m ← m+1
//If it does not exist, remove the current child node
return  ChooseNode.Nodes
End.
```

Through the selection structure algorithm, the query string in the selection node is split to obtain the database table name, column name and query conditions, so that the database can obtain the selection parameters. Then select parameters to filter nodes and complete the structure change of node tree template. This method is to select nodes, sentence selection structure is to select sentence nodes, and phrase selection structure is to select phrase nodes. Both structures can be realized through this method.

Compared with the replacement of Word document template, this method abstracts each node into a class, including methods and attributes, and no longer relies on external engines to modify the template structure. In the selection structure, the method and attribute in the node class are selected to obtain the data in the database, and then the automatic selection of sentences and texts is completed through the data to achieve the goal of data-driven, so as to achieve the adaptability of the method. The template tree is controlled through data. To modify the template structure, only the node parameters need to be modified, which improves the flexibility of the template and reduces the cost of later modification and maintenance.

4. Application examples

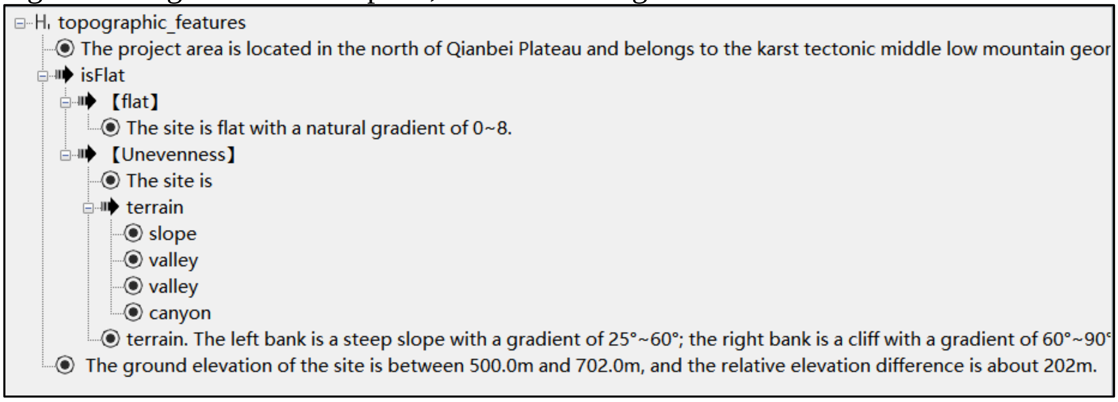
The engineering geological survey report is the final result of the survey and an important basis for relevant engineering design and construction [9]. The engineering geological survey report is a normative document, which shall comply with relevant geological survey requirements and specifications. A paragraph in a geological survey report is used as a template. The paragraph contains two selection structures: a phrase selection structure; A statement selects the structure, as shown in the Figure 2.



The project area is located in the north of Qianbei Plateau and belongs to the karst tectonic middle low mountain geomorphic unit. [The site is flat with a natural gradient of 0~8.] [The site is<slope | valley | valley | canyon>terrain. The left bank is a steep slope with a gradient of 25° ~60° ; the right bank is a cliff with a gradient of 60° ~90° .] The ground elevation of the site is between 500.0m and 702.0m, and the relative elevation difference is about 202m.

Figure 2: Word original template of a geological survey report section.

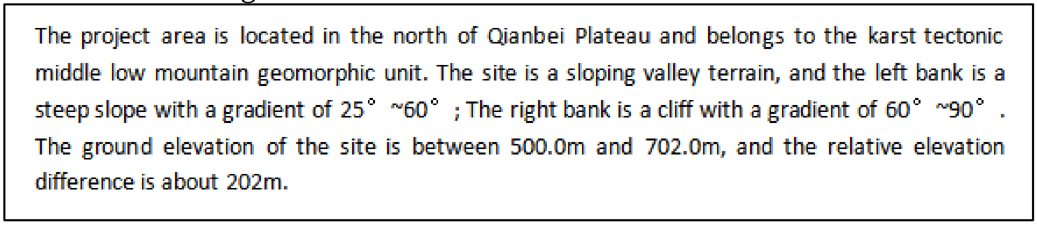
Sentence selection refers to the selection of flat and uneven terrain in a paragraph. It is surrounded by "[]" in the figure. Each surrounded sentence represents an option; Text selection refers to the selection of different site topographies in the site. It is surrounded by "<>" symbols in the figure, and each option is separated by "|".The following node tree templates are made according to the original Word template, as shown in Figure 3.



```
H1 topographic_features
  The project area is located in the north of Qianbei Plateau and belongs to the karst tectonic middle low mountain geor
  isFlat
    【flat】
      The site is flat with a natural gradient of 0~8.
    【Unevenness】
      The site is
      terrain
        slope
        valley
        valley
        canyon
      terrain. The left bank is a steep slope with a gradient of 25°~60°; the right bank is a cliff with a gradient of 60°~90°
  The ground elevation of the site is between 500.0m and 702.0m, and the relative elevation difference is about 202m.
```

Figure 3: Node tree template of a geological survey report section.

In this template, isFlat and terrain are two selection nodes. isFlat controls whether the terrain is flat. Choose one of the two options: flat and uneven; Terrain controls the selection of site terrain. Select one from slope, valley, valley and gorge. When the selection parameters obtained by two selection nodes through the query string are respectively uneven and valley, the paragraph results obtained are shown in the Figure 4.



The project area is located in the north of Qianbei Plateau and belongs to the karst tectonic middle low mountain geomorphic unit. The site is a sloping valley terrain, and the left bank is a steep slope with a gradient of 25° ~60° ; The right bank is a cliff with a gradient of 60° ~90° . The ground elevation of the site is between 500.0m and 702.0m, and the relative elevation difference is about 202m.

Figure 4: Target document of a geological survey report section.

The above examples show that this method can complete the generation of normative document paragraphs, and it is better in paragraphs with multiple choice structures.

5. Conclusion

In order to solve the problem that the replacement method of Word template can not deal with the selection structure well in document generation, this paper proposes an adaptive selection algorithm of paragraph selection structure based on complex data types. This method completes the modification of the template structure by abstracting nodes into classes and encapsulating methods and attributes of classes. The results show that the method is simple and applicable in solving the problem of structure selection. This method improves the efficiency of automatic generation of standard documents and reduces the proportion of manual operations.

Acknowledgements

The author thanks Jian Lin and Zhinan Lin for their help in the research process.

References

[1] Li Xueqing, Wang Shi, Wang Zhujun, Zhu Junwu. Overview of Natural Language Generation [J]. Computer Application, 2021, 41 (05): 1227-1235.

[2] Bai Xinyu Design and Implementation of Digestive Endoscopy Report Automatic Generation System Based on Intelligent Template [D]. Shandong University, 2021. DOI: 10.27272/d.cnki.gshdu. 2021.004548.

[3] Wang Xiliang. Research and Application of Public Meteorological Service Document Automation [J]. Middle and Low Latitude Mountain Meteorology, 2018, 42 (03): 95-98.

[4] Ma Yongzhi, Zhang Jipeng, Zheng Yihua, Wang Dechang, Liu Fuwang, Cao Zhaobin. Research on the automatic generation of the documents of the automobile radiator design calculation platform [J]. Journal of Qingdao University (Engineering Technology Edition), 2013, 28 (04): 71-75. DOI: 10.13306/j.1006-9798.2013.04.016.

[5] Fu Yan, Ji Min, Jia Ning. Design and Implementation of Oil Fingerprint Identification Report Generation System [J]. Ocean Information, 2018, 33 (02): 58-62. DOI: 10.19661/j.cnki.mi. 2018.02.011.

[6] Jiang Peng, Xu Feng, Qi Rongzhi. Building an intelligent generation model of flood control documents based on cloud platform [J]. Water Resources Informatization, 2013 (03): 25-32. DOI: 10.19364/j.1674-9405.2013.03.008.

[7] Wang Zhengmin, Zhang Taihong, Li Yongke, etc. FreeMarker template engine online dynamic generation of Excel and Word documents [J]. Computer and Modernization, 2016 (4): 109-113.

[8] Luo Rong, Huang Jun, Li Maofeng, Liu Zhiqin. A Fast Generation Method of Complex Documents Based on Word Template [J]. Computer Application and Software, 2020, 37 (10): 57-63.

[9] Du Juan. Preparation of Engineering Geological Survey Report [J]. Chinese and Foreign Entrepreneurs, 2015 (12): 219-220.