ISSN 2616-5775 Vol. 7, Issue 5: 109-115, DOI: 10.25236/AJCIS.2024.070514

Automatic Generation Method for Tables with Merged Cells

Maolong Teng[1,a], Zhinan Lin[2,b,*], Chaoyue Liu[3,c]

[1]*Hunan University of Science and Technology, Xiangtan, China*
[2]*Hunan Vocational Institute of Technology, Xiangtan, China*
[3]*Guizhou Provincial Transportation Planning Survey and Design Research Institute Co., Ltd., Guiyang, China*
[a]*1929931622@qq.com,* [b]*409435402@qq.com,* [c]*464683877@qq.com*
[*]*Corresponding author*

Abstract: *This paper proposes a complex table generation method based on a multi tree structure to address the problems of large data volume, repetitive and cumbersome processes, and inability to automatically adjust the structure according to data content in the current automatic table generation process. This method uses a tree structure to represent data tables, categorizes the nodes in the tree structure in detail, and sets corresponding properties and methods according to different generation requirements. This tree structure is defined as a multi-dimensional relationship oriented table structure tree. In the process of automatic table generation, a method for automatic data query and a method for calculating the number of rows and columns that need to be merged based on the correlation between data are designed. This method improves the efficiency and accuracy of table making, and can effectively adapt to the diversity and complexity of table structures. It can not only adapt to different scenarios and requirements of table generation tasks, but also improve the efficiency and accuracy of table generation.*

Keywords: *Automatically generate tables, Cell merging, Related relationships*

1. Introduction

A table is a compact, well formatted, and easy to visually display form of data, widely used in various industries. Tables have excellent data organization and expression capabilities, making it easy to transfer and exchange various types of data such as text and numbers [1], It is the underlying data organization form of relational databases and various information processing systems, where a large amount of information data is organized and represented in the form of tables.

Automatic table generation refers to the process of using computer programs to automatically create, fill in, and format tables [2]. The technology of automatic table generation has been applied in multiple fields, and in the fields of business and finance, many complex data need to be organized, analyzed, and reported. The tables used by domestic enterprises have great complexity because they require the layout of internal header cells to express rich semantic information of data [3]. Each cell's data has relatively independent and rich semantics, and it is necessary to control and process each cell as an independent element. These cells are cleverly arranged in a hierarchical nested combination form within the table to express deeper data semantics. In order to better manage and analyze data, Li Jinhao [4] used the programming language C # to automatically generate tables, combined with SQL statements to filter and export data from the data integration platform and write them into the table. With the breakthrough of natural language technology, ChatGPT has been widely popular recently. Using ChatGPT can generate table style data, but after pasting it into Word, it needs to be reformatted, and the data in it also comes from the network. ChatExcel can achieve complex operations in Excel through dialogue and supports one click export into standard Excel tables. However, it needs to upload the original basic table as the data source, and cannot directly generate and modify the target table from the data in the database.

In response to the above issues, this article proposes a new method for automatic table generation. After binding the data source, it can automatically query data and calculate the number of rows and columns that need to be merged based on the entity properties of the table. This allows for flexible adjustments to the generated table according to specific data content and display requirements.

ISSN 2616-5775 Vol. 7, Issue 5: 109-115, DOI: 10.25236/AJCIS.2024.070514

2. Basic Content of the Table

Tables are a structured form used for organizing, presenting, and comparing data, widely used in business, academia, science, and daily life. A table is a form of data organization used for collecting, organizing, organizing, and analyzing data [5]. Cells are the basic elements that make up a table, which can contain text, numbers, and other content. The rows and columns of a table are independent and interrelated, and by combining and arranging different cells, the relationships and patterns between data can be clearly displayed.

v_{11}	v_{12}	…	v_{1y}
v_{21}	v_{22}	…	v_{2y}
…	…	…	…
v_{x1}	v_{x2}	…	v_{xy}

Figure 1: Sample form.

The structure in the example table in Figure 1 will be used to illustrate some concepts, and the basic definition of the content in the table is as follows:

Definition 1 (Table). The table consists of an ordered set of x rows and y columns. The table is divided into two areas: indicator field and data field. The indicator field is located above the table, and the data field is located below the indicator field.

Definition 2 (Cells). Intersections in a table, where each cell is used to store specific data or information. Each cell can contain different types of data, such as text, numbers, dates, etc. The cells located in the indicator field are the attribute list cells, and the cells located in the data field are the attribute value cells.

Definition 3 (Single valued data and multi valued data). As shown in Figure 1, each intersection between rows and columns determines cell c_{ij}, which has a value of v_{ij}, where $1 \leq i \leq x$ and $1 \leq j \leq y$. The value v_{ij} comes from a set of free labels $L = \{L_1, \dots, L_y\}$ or from a set of data $D = \{D_1, \dots, D_y\}$. If the value v_{ij} comes from a certain data value in the set, it is called single valued data. If the value v_{ij} is composed of a set of data values, it is called multi valued data.

Definition 4 (Continuous Cells). Let c_{ab} and c_{de} be the two cells in the table, where a and d are row indexes and b and e are column indexes. If d=a+1 or e=b+1, then c ab and c de are called continuous cells.

Let m, n, and p be three arbitrary cells. If m and n are continuous cells, and n and p are continuous cells, then m, n, and p form a set of continuous cells.

Definition 5 (Merge Cells). Let mc be a group of consecutive cells in a table. If all of their elements are associated with the same value v, then mc is called a merged cell, where $v \in L$ or $v \in D$. It should be noted that only continuous cells can form merged cells.

3. Multidimensional Relation Oriented Table Structure Tree

When generating tables, a tree structure is used to represent the table. Based on the data sources and relationships between data in the table, the nodes in the structure tree are classified in detail and corresponding properties and methods are set according to different generation requirements. This tree structure is defined as a multi-dimensional relationship oriented table structure tree, which can fully represent the table structure. Through relevant methods, automatic data processing can be achieved to generate the target table.

3.1. Node Classification

In order to better meet the generation needs of various tables, more detailed classification of nodes in the tree has been carried out. At different stages of table generation, the content of nodes in the tree also changes accordingly. Firstly, when creating a table template and setting up the association relationships and data sources between the table indicator fields and data field cells, a table template tree is constructed. The nodes in the tree are all table template nodes used to describe the structure and association

ISSN 2616-5775 Vol. 7, Issue 5: 109-115, DOI: 10.25236/AJCIS.2024.070514

relationships of the table. The node types used in building a table template tree vary depending on the various combinations of tables. Once the structure of the table, data content, and the number of rows and columns to merge cells are determined, a new table object tree needs to be created, which is generated from the table template tree. In the table object tree, all nodes in the tree are table object nodes used to store specific data information for generating instance tables. Different instance tables can be generated according to requirements to meet specific data presentation or analysis requirements. The node classification of the multi-dimensional relationship oriented table structure tree is shown in Figure 2.

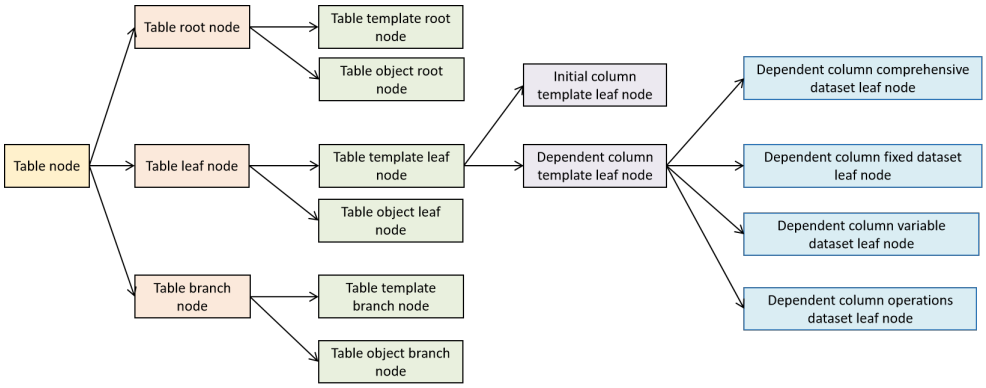


Figure 2: Multi-dimensional relationships guide common node types in a table structure tree.

3.2. Node UML Model

After classifying the nodes in the multi-dimensional relationship oriented table structure tree, the attributes and methods contained in each node in the table template tree and table object tree are clarified, which helps to ensure that each node plays its due role in the model.

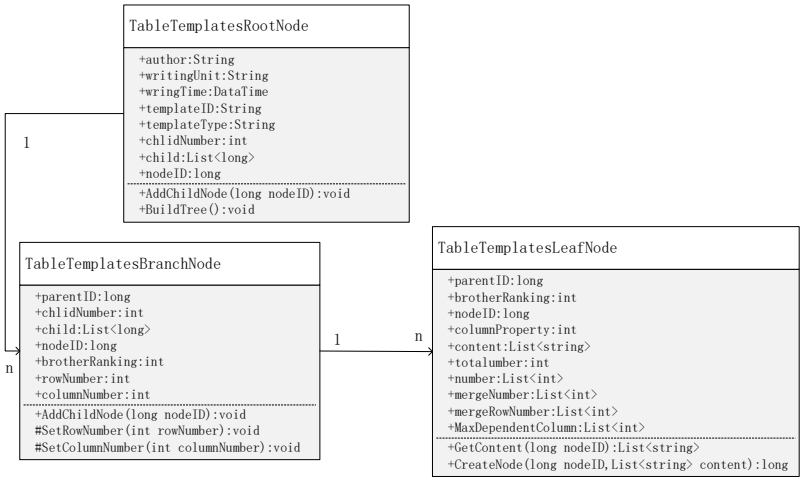


Figure 3: Table Object Tree Entity Relationship Diagram.

The root node of the table template in the table template tree corresponds to multiple table template branch nodes. Table template branch nodes can handle the content of the table indicator field well, store the content of each cell in the indicator field in the node, and the tree structure can represent the structure of the table indicator field well. At the same time, according to the node hierarchy and corresponding situation, the number of rows and columns that need to be merged for each cell can be calculated and stored in the node, which is convenient for generating instance tables later. Table template leaf nodes indicate the content of the attribute value cell array in the column data field. The association between table template leaf nodes can represent the relationship between cells in each column well. By setting the properties and methods of nodes, the data source and calculation method of the table can be clarified. Can query data according to settings, and set the number of rows and columns that need to be merged for each cell based on the correspondence between data. Table template leaf nodes are divided into initial column template leaf nodes and subordinate column template leaf nodes based on the type of affiliation. The subordinate column template leaf nodes are further divided into subordinate column fixed dataset leaf nodes based on the form and processing method of the query data Dependent column variable dataset leaf node, Dependent column comprehensive dataset leaf node, and Dependent column operation dataset

ISSN 2616-5775 Vol. 7, Issue 5: 109-115, DOI: 10.25236/AJCIS.2024.070514

leaf node.

The table object tree stores the data in the instance table, including the data content in cells and the number of rows and columns occupied by cells. It is generated by the table template tree by querying data and calculating the number of rows and columns that need to be merged in cells. At different time periods, due to the update of database content data, the content in the generated table object tree may vary, and the generated table instances may also be different. The table object tree contains three types of nodes, namely the table object root node, table object branch node, and table object leaf node. The node properties and methods are shown in Figure 3. The table object tree can represent the content and structure of the instance table in a concise and clear manner.

4. Automatic Table Generation Method

In the process of generating a table, it is necessary to traverse the table template tree to generate a table object tree. The data content of each cell, the number of rows and columns occupied by each cell, and other information are stored in the table object tree. Then, the table object tree generates an instance table. In the process from setting the template tree to generating the table object tree, it is necessary to determine the content of the table indicator field, calculate the number of rows and columns occupied by each attribute list cell in the indicator field, and search for the content of attribute value cells in the data field based on the set data source query field. According to the query content and corresponding relationship, calculate the number of rows and columns occupied by each attribute value cell in the data field.

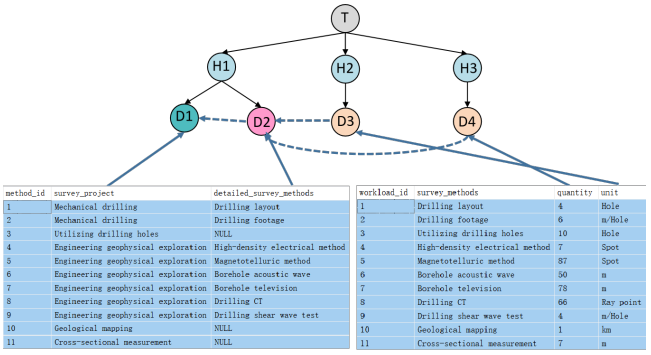
4.1. Automatic Data Query Method

The data in the query table is a crucial and indispensable part of the table generation process. Table data is the foundation of the table's existence, and it presents various information in a structured manner. The data sources for tables are diverse, including databases, data files, user input, or content from other systems. At the same time, some data in the table needs to be processed and obtained through computer processing. In order to automatically obtain the dataset, data sources and query conditions corresponding to the columns were added to the leaf nodes of the table template. Based on the relevant fields and associated conditions of the added data query, I converted them into a Select statement in SQL for data queries. The Select statement has flexible usage, and its query example is shown in Figure 4.

```
select rock_mass_information.rock_masss_name from
rock_mass_information,test_paragraphs_information where
rock_mass_information.paragraph_name=test_paragraphs_information.paragraph_name
```

Figure 4: SQL statement instance.

In the data query process, the first step is to query the data content in the initial column of the table, because the data content in the initial column of the table does not depend on other columns. Then, the data content of other subordinate columns is queried. During the query process, the content of the subordinate columns needs to be queried based on the content of the main column cells. The number of queries for each main column cell is stored in an array, and this related information is saved in the table template leaf node to prepare for the number of rows that need to be merged in the subsequent cells. The data content in the database table stored in each leaf node of the instance table template tree is shown in Figure 5.



method_id	survey_project	detailed_survey_methods
1	Mechanical drilling	Drilling layout
2	Mechanical drilling	Drilling footage
3	Utilizing drilling holes	NULL
4	Engineering geophysical exploration	High-density electrical method
5	Engineering geophysical exploration	Magnetotelluric method
6	Engineering geophysical exploration	Borehole acoustic wave
7	Engineering geophysical exploration	Borehole television
8	Engineering geophysical exploration	Drilling CT
9	Engineering geophysical exploration	Drilling shear wave test
10	Geological mapping	NULL
11	Cross-sectional measurement	NULL

workload_id	survey_methods	quantity	unit
1	Drilling layout	4	Hole
2	Drilling footage	6	m/Hole
3	Utilizing drilling holes	10	Hole
4	High-density electrical method	7	Spot
5	Magnetotelluric method	87	Spot
6	Borehole acoustic wave	50	m
7	Borehole television	78	m
8	Drilling CT	66	Ray point
9	Drilling shear wave test	4	m/Hole
10	Geological mapping	1	km
11	Cross-sectional measurement	7	m

Figure 5: Determine the query content of the instance table.

ISSN 2616-5775 Vol. 7, Issue 5: 109-115, DOI: 10.25236/AJCIS.2024.070514

4.2. Cell Calculation Method

After clarifying the structure of the table and the relationships between cells, it is necessary to fill the data into the cells. However, in tables containing different data, the relationships between the columns of the data are different, and the dependency relationships between the column cells affect the display of the content in the data field below. The attribute value cells in the main column will merge cells based on their corresponding relationships and specific data content. Automatic merging of attribute value cells is a challenge in the table generation process, and how to calculate the number of rows and columns that need to be merged for each attribute value cell is the key. Taking Table 1 as an example, the physical workload summary table will be used to calculate the number of rows and columns for cell merging.

Table 1: Summary table of physical workload.

Project		Unit	Quantity
Mechanical drilling	Drilling layout	Hole	4
	Drilling footage	m/Hole	6
Utilizing drilling holes		Hole	7
Engineering geophysical exploration	High-density electrical method	Spot	10
	Magnetotelluric method	Spot	87
	Borehole acoustic wave	m	50
	Borehole television	m	78
	Drilling CT	Ray point	66
	Drilling shear wave test	m/Hole	4
Geological mapping		km	1
Cross-sectional measurement		m	7

4.2.1. Calculate the Number of Merged Rows in Cells

After determining the content and number of datasets corresponding to each column and leaf node in the table tree, it is also necessary to determine the number of rows and columns that need to be merged in the cells based on logical relationships. For calculating the number of merged rows, the main parameters used are rowNumber, mergeNumber, and mergeRowNumber. After executing SQL statements and querying data, the contents of the rowNumber, mergeNumber, and mergeRowNumber arrays in the leaf nodes of the table template are shown in Table 2. Because there are multiple columns under the same attribute list cell in this table, the data in the second column of the table depends on the data in the first column. Therefore, when the second column finds empty results based on the data in the first column, it is necessary to copy the content of its corresponding main column cell to ensure that the third and fourth columns can correspond to the first and second columns when searching for data. At the same time, the number of rows occupied by the data will be recorded as 1, as it has already copied the content of its corresponding main column cell.

Table 2: Calculate the content of the front row array in the physical workload summary table.

Leaf node	rowNumber array	mergeNumber array	mergeRowNumber array
D1	1,1,1,1,1	1,1,1,1,1	1,1,1,1,1
D2	1,1,1,1,1,1,1,1,1,1	2,1,7,1,1	2,1,7,1,1
D3	1,1,1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1
D4	1,1,1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1

Due to the existence of only one to many relationships in the test paragraph result analysis table, only the cells in the first column need to be merged. After updating, the contents of each array are shown in Table 3.

Table 3: The content of the row array after calculating the physical workload summary table.

Leaf node	rowNumber array	mergeNumber array	mergeRowNumber array
D1	2,1,7,1,1	2,1,7,1,1	2,1,7,1,1
D2	1,1,1,1,1,1,1,1,1,1	2,1,7,1,1	2,1,7,1,1
D3	1,1,1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1
D4	1,1,1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1

After calculation, the rowNumber array in each leaf node of the template tree of the physical workload summary table represents the number of rows occupied by each attribute value cell from top to bottom in the corresponding column's data field.

ISSN 2616-5775 Vol. 7, Issue 5: 109-115, DOI: 10.25236/AJCIS.2024.070514

4.2.2. Calculate the Number of Columns for Cell Merging

In some tables, there is a situation where multiple columns share the same attribute list cell, so it is necessary to determine whether the attribute list cell needs to cross columns and the number of columns it crosses based on the data content. The main parameters that need to be used are columnNumber and virtualValue. Taking the physical workload summary table as an example, when conducting data queries, the number of rows that need to be merged for each attribute value cell corresponding to the main column attribute value cell is calculated based on the query results. Only when a cell in the dependent column cannot find data, the corresponding main column cell needs to be merged across rows. The related array data content is shown in Table 4.

Table 4: The content of the array before calculating the physical workload summary table.

Leaf node	columnNumber array	virtualValueNumber array
D1	1,1,1,1,1	0,0,0,0,0
D2	1,1,1,1,1,1,1,1,1,1	0,1,0,1,1
D4	1,1,1,1,1,1,1,1,1,1	0,0,0,0,0
D4	1,1,1,1,1,1,1,1,1,1	0,0,0,0,0

When calculating, it is necessary to traverse the leaf nodes of the table template, add the virtualValue Number array from the dependent columns to the columnNumber array of the main column, and the resulting columnNumber array is the number of columns that need to be merged from top to bottom attribute value cells in each column data field. The calculated data content is shown in Table 5.

Table 5: The content of the column array after calculating the physical workload summary table.

Leaf node	columnNumber array	virtualValueNumber array
D1	1,2,1,2,1	0,0,0,0,0
D2	1,1,1,1,1,1,1,1,1,1	0,1,0,1,1
D4	1,1,1,1,1,1,1,1,1,1	0,0,0,0,0
D4	1,1,1,1,1,1,1,1,1,1	0,0,0,0,0

4.3. Cell Determination Process

This section will introduce the process of determining the number of rows and columns when merging table cells. Figure 6 is a display of the data in each leaf node and its corresponding relationship in a table template. From this figure, it can be seen that this table contains multiple levels of master-slave relationships. The data in T2 and T3 are related to the data in T1, T4 is related to the data in T3, T5 is related to the data in T4, and after determining the data content and correlation relationship, it is necessary to calculate the number of rows and columns that need to be merged for each cell in each column.

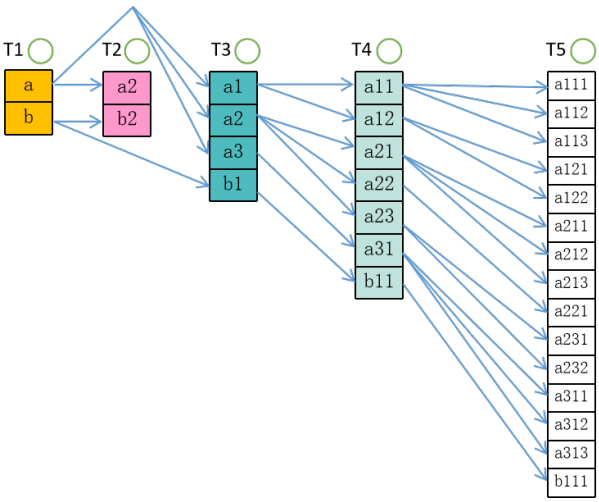


Figure 6: Data examples in each leaf of a table template.

Figure 7 shows the number of rows and columns of cells in the column determined after each processing step of the data in Figure 6. At the beginning, the default number of rows for each cell is 1. When traversing leaf node T2, the data in T2 and T1 have a one-to-one relationship, so the number of rows in cells a, b, a2, and b2 is 1. When traversing leaf node T3, based on the data in T3, it can be

ISSN 2616-5775 Vol. 7, Issue 5: 109-115, DOI: 10.25236/AJCIS.2024.070514

determined that a in the first column corresponds to a1, a2, a3, and b pairs b1. At this time, the number of rows in each cell in the third column is 1, so the number of rows crossed by a in the first column is 3. The number of rows is 1. When traversing the leaf node T4, based on the data in T4, the corresponding number of rows for a1, a2, a3, and b1 in the third column can be determined to be 2, 3, 1, and 1. After modifying the third column, the number of rows merged in the first column needs to be updated again from the data in the third column, as it is related to the number of rows in the third column cells. When traversing the leaf node T5, like T4, the corresponding number of rows in the main column needs to be determined forward in sequence based on the master-slave relationship, At this point, according to the method of calculating the number of merged rows in the main column cells based on the content of the dependent columns in the first part of the cell calculation method, the data of the cells in nodes T1, T3, T4, and T5 have been determined. Then, the second part of the cell calculation method needs to modify the content of node T2 by calculating the number of merged rows in the dependent column cells based on the content of the main column. As the T2 node and T1 node have a one-to-one relationship, the number of rows in the cells in node T1 has been modified by the content of other dependent columns, so it is necessary to correspond the number of rows in the cells in node T2 to the number of rows in the cells in node T1. Due to the absence of multiple columns sharing the same attribute list cell in this table, there is no need for cells to cross columns. Therefore, the number of columns occupied by each cell is 1. At this point, the calculation of cells has been completed, and the number of rows and columns in the attribute value cells corresponding to each leaf node in the table template tree has been determined. The instance table can be generated later.

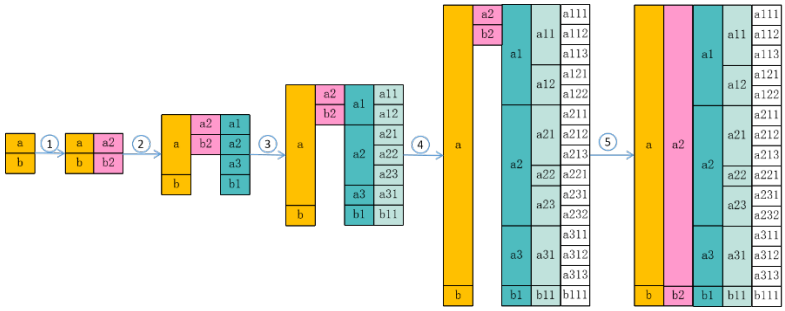


Figure 7: The process of determining the number of cell rows per column.

5. Conclusions

This article elaborates on the key steps and methods in the process of table generation, ensuring the clarity and logicality of the table structure by accurately calculating the number of rows and columns that need to be merged into cells in the indicator field. This step is crucial as it directly affects the overall layout and readability of the table. Implemented automatic data querying and conversion, ensuring that the information extracted from the data source can be accurately and accurately filled into the table. This not only improves the efficiency of data processing, but also avoids errors caused by manual input. When generating data field cells, determine the number of rows and columns to merge based on the master-slave relationship between the data. This method can accurately reflect the hierarchy and correlation between data, making the generated tables more targeted and practical. By integrating these technologies, efficient and accurate table generation has been achieved, and the table structure can be dynamically adjusted according to changes in data.

References

[1] Hu J, Kashi R S, Lopresti D P, et al. Medium-independent table detection[C]//Document Recognition and Retrieval VII. SPIE, 1999, 3967: 291-302.
[2] Aithal S G, Rao A B, Singh S. Automatic question-answer pairs generation and question similarity mechanism in question answering system[J]. Applied Intelligence, 2021: 1-14.
[3] Gottschalk S, Demidova E. Tab2KG: Semantic table interpretation with lightweight semantic profiles [J]. Semantic Web, 2022, 13(3): 571-597.
[4] Li Jinhao. Research and practice of report automatic generation technology based on C # [J]. Modern Computer, 2020, (36): 95-99
[5] Zehra S, Mohsin S F M, Wasi S, et al. Financial knowledge graph based financial report query system [J]. IEEE Access, 2021, 9: 69766-69782.