ISCAS 中国科学院软件研究所学术年会'2021
暨计算机科学国家重点实验室开放周

电子表格语义结构识别

张雅坤，崔紫玉，吕潇，董浩宇，窦文生，韩石，张冬梅，魏峻，叶丹

Semantic Table Structure Identification in Spreadsheets

The 30th of ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA'21)

联系方式：窦文生， wensheng@iscas.ac.cn

Table Structure

☐**Complex and informal structures within spreadsheet tables.**

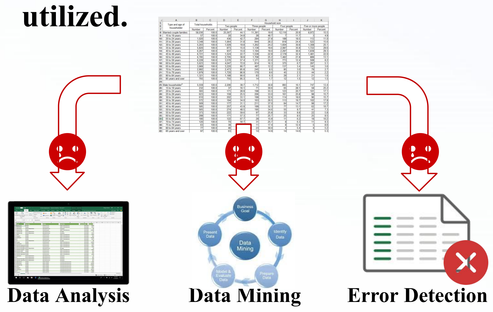
Parent-child relation

Sibling relation

	A	B	C	D	E	F
1			Month			
2	Departme	Country	January	February	March	Total
3			Cost	Cost	Cost	Cost
4	Sales	America	376,589	315,644	453,243	=SUM(C4:E4)
5		Japan	557,890	265,789	254,565	=SUM(C5:E5)
6		Total	=SUM(C4:C5)	581,433	707,808	=SUM(C6:E6)
7	Market	America	654,234	485,432	346,218	=SUM(C7:E7)
8		Japan	234,200	252,432	433,689	=SUM(C8:E8)
9		England	23,455	161,132	153,368	=SUM(C9:D9)
10		Total	=SUM(C6:C9)	=SUM(D6:D9)	=SUM(E6:E9)	=SUM(C10:E10)
11	*The costs of Sales and Market department.					

No Clear Table Structures

☐**Spreadsheet data cannot be fully utilized.**



Semantic Table Structure

☐**Divide header types according to their purpose.**

- Index
- Value name
- Index name
- Aggregation

	A	B	C	D	E	F
1			Month			
2	Department	Country	January	February	March	Total
3			Cost	Cost	Cost	Cost
4	Sales	America	376,589	315,644	453,243	=SUM(C4:E4)
5		Japan	557,890	265,789	254,565	=SUM(C5:E5)
6		Total	=SUM(C4:C5)	581,433	707,808	=SUM(C6:E6)
7	Market	America	654,234	485,432	346,218	=SUM(C7:E7)
8		Japan	234,200	252,432	433,689	=SUM(C8:E8)
9		England	23,455	161,132	153,368	=SUM(C9:D9)
10		Total	=SUM(C6:C9)	=SUM(D6:D9)	=SUM(E6:E9)	=SUM(C10:E10)
11	*The costs of Sales and Market department.					

Index

Semantic Table Structure

☐**There exist certain relations among headers.**

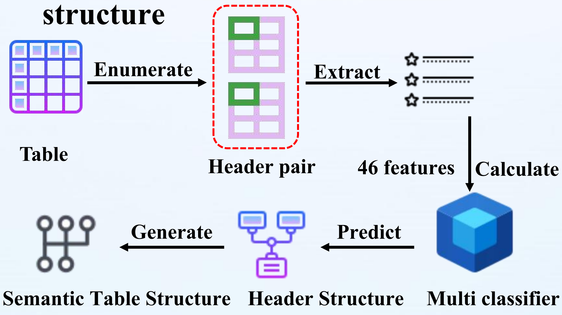
- Parent - child
- Aggregation relation
- Sibling relation
- Index name - index

	A	B	C	D	E	F
1			Month			
2	Department	Country	January	February	March	Total
3			Cost	Cost	Cost	Cost
4	Sales	America	376,589	315,644	453,243	=SUM(C4:E4)
5		Japan	557,890	265,789	254,565	=SUM(C5:E5)
6		Total	=SUM(C4:C5)	581,433	707,808	=SUM(C6:E6)
7	Market	America	654,234	485,432	346,218	=SUM(C7:E7)
8		Japan	234,200	252,432	433,689	=SUM(C8:E8)
9		England	23,455	161,132	153,368	=SUM(C9:D9)
10		Total	=SUM(C6:C9)	=SUM(D6:D9)	=SUM(E6:E9)	=SUM(C10:E10)
11	*The costs of Sales and Market department.					

Parent - child

Tasi

☐**Automatically identify semantic structure**



Tasi Evaluation

☐**The accuracy of semantic table structure identification is 44.1%.**

339/625=54.2%

	A	B	C	D	E	F
1			Month			
2	Department	Country	January	February	March	Total
3			Cost	Cost	Cost	Cost
4	Sales	America	376,589	315,644	453,243	=SUM(C4:E4)
5		Japan	557,890	265,789	254,565	=SUM(C5:E5)
6		Total	=SUM(C4:C5)	581,433	707,808	=SUM(C6:E6)
7	Market	America	654,234	485,432	346,218	=SUM(C7:E7)
8		Japan	234,200	252,432	433,689	=SUM(C8:E8)
9		England	23,455	161,132	153,368	=SUM(C9:D9)
10		Total	=SUM(C6:C9)	=SUM(D6:D9)	=SUM(E6:E9)	=SUM(C10:E10)
11	*The costs of Sales and Market department.					

363/518=70.1%

282/639=44.1%

TasiError

☐**Structure-based spreadsheet error detection.**

Index set

	A	B	C	D	E	F
1			Month			
2	Department	Country	January	February	March	Total
3			Cost	Cost	Cost	Cost
4	Sales	America	376,589	315,644	453,243	=SUM(C4:E4)
5		Japan	557,890	265,789	254,565	=SUM(C5:E5)
6		Total	=SUM(C4:C5)	581,433	707,808	=SUM(C6:E6)
7	Market	America	654,234	485,432	346,218	=SUM(C7:E7)
8		Japan	234	252,432	433,689	=SUM(C8:E8)
9		England	5	161,132	153,368	=SUM(C9:D9)
10		Total	C6:C9)	=SUM(D	UM(E6:E9)	=SUM(C10:E10)
11	ent.					

Formula pattern is SUM(R[-2]C:R[-1]C)

Missing formula: 581,433=315,644+265,789

TasiError Evaluation

☐**TasiError outperforms existing tools (Precision：75.2%，Recall：82.9).**