



电子表格中可扩展单元组识别

窦文生, 韩石, 徐良, 张冬梅, 魏峻

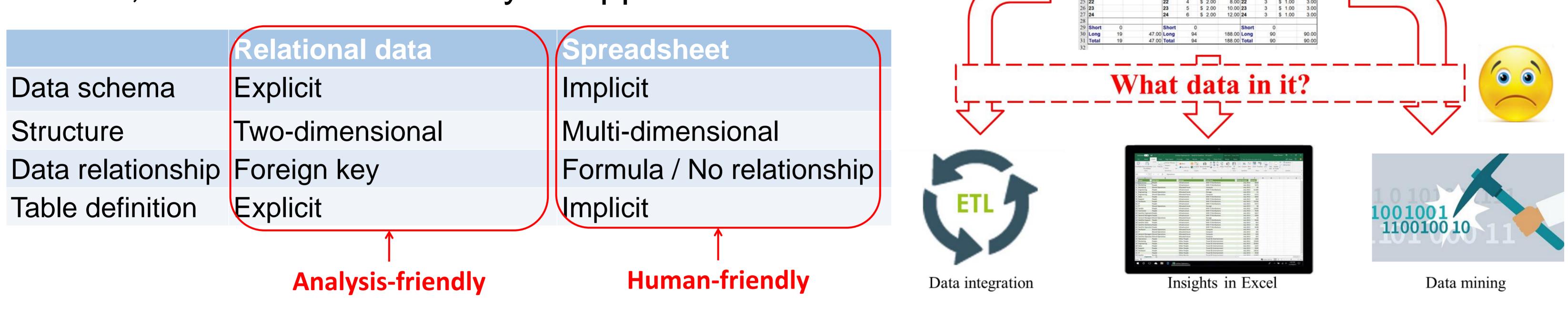
Expandable Group Identification in Spreadsheets, ASE 2018, pp. 498-508. 联系方式: 窦文生, wensheng@iscas.ac.cn, http://www.tcse.cn/~wsdou

Spreadsheet structure information is missing

Spreadsheets are designed to be **interpreted by**



human, rather than data analysis applications.



Expandable group

Expandable group: A basic unit and all its repeating units form an expandable group.

The corresponding cells in an expandable group have > Based on identified expandable groups, we can the similar formats and semantics, e.g., cells [A2:D31]
The corresponding cells in an expandable group have > Based on identified expandable groups, we can transform spreadsheet data into relational data,



	Α	В	С	D	F	F		C	Н	Ι	J		K	L	
1															
2	New York					Chicago					Atlanta				
3	Hour	Volume	Price	Position	Hour	Volume	Р	rice	Position	Hour	Volume	F	Price	Position	
4	1	1	\$ 2.00	2.00	1	1	\$	2.00	2.00	1	2	\$	1.00	2.00	
5	2	2	\$ 3.00	6.00	2	2	\$	2.00	4.00	2	4	\$	1.00	4.00	
6	3	3	\$ 1.00	3.00	3	4	\$	2.00	8.00	3	5	\$	1.00	5.00	
7	4	3	\$ 1.00	3.00	4	5	\$	2.00	10.00	4	6	\$	1.00	6.00	
8	5	5	\$ 5.00	25.00	5	6	\$	2.00	12.00	5	7	\$	1.00	7.00	
23	20				20	2	\$	2.00	4.00	20	3	\$	1.00	3.00	
24	21				21	3	\$	2.00	6.00	21	3	\$	1.00	3.00	
25	22				22	4	\$	2.00	8.00	22	3	\$	1.00	3.00	
26	23				23	5	\$	2.00	10.00	23	3	\$	1.00	3.00	
27	24				24	6	\$	2.00	12.00	24	3	\$	1.00	3.00	
28															
29	Short	0			Short	0				Short	0				
30	Long	19		47.00	Long	94			188.00	Long	90			90.00	
31	Total	19		47.00	Total	94			188.00	Total	90			90.00	
32					•		•	-	/				-		

which can be used by other data analysis tools.

City	Hour	Volume	Price	Position	
New York	1	1	2.00	2.00	
New York	2	2	3.00	6.00	
New York	3	3	1.00	3.00	
Chicago	1	1	2.00	2.00	
Chicago	2	2 6	2.00	4.00	
Chicago	3	4	2.00	8.00	
Atlanta	1	Analysis-	2.00		
Atlanta	2	4	1.00	4.00	
Atlanta	3	5	1.00	5.00	

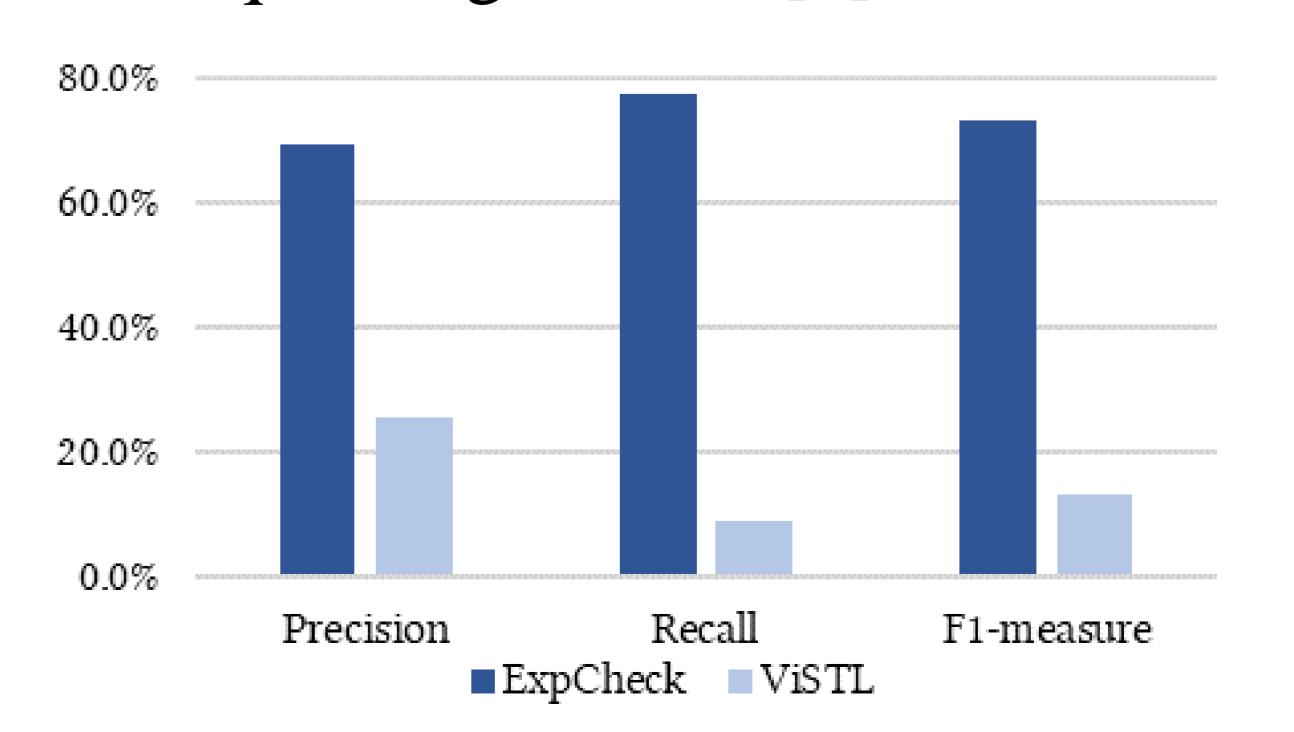
ExpCheck

Key observation: Units in an expandable group have the similar formats and semantics.

Evaluation

ExpCheck performs much better than existing techniques, e.g., ViSTL[1].

	Α	В	С	D	Е	F		G	Η	Ι	J	Κ	L	
1	Balancing													
2	New York				Chicago					Atlanta				
3	Hour	Volume	Price	Position	Hour	Volume	P	rice	Position	Hour	Volume	Price	Position	
4	1	1	2.00	2.00	1	1	¢	2 00	200	4	2	¢ 1.00	2.00	
5	2	2	3.00	6.00	2	2	S	imi	lar se	ema	ntics	1.00	4.00	
6	3	3	\$ 1.00	3.00	3	4	¥	2.00	0.00	-		1.00	5.00	
7	4	3	\$ 1.00	3.00	4	5	\$	2.00	10.00	4	6	\$ 1.00	6.00	
8	5	5	\$ 5.00	25.00	5	6	\$	2.00	12.00	5	7	\$ 1.00	7.00	
23	20		K		20	2	\$	2.00	4.00	20	3	\$ 1.00	3.00	
24	21				21	3	\$	2.00	6.00	21	3	\$ 1.00	3.00	
25	22				24	1	¢	2.00	8.00	22	3	\$ 1.00	3.00	
26	23		San	าe fo	rm	ats		2.00	10.00	23	3	\$ 1.00	3.00	
27	24					2.00	12.00	24	3	\$ 1.00	3.00			
28														
29	Short	0			Short	0				Short	0			
30	Long	19		47.00	Long	94			188.00	Long	90		90.00	
31	Total	19		47.00	Total	94			188.00	Total	90		90.00	



[1] R. Abraham et al., Inferring Templates from Spreadsheets, ICSE 2006.