Integrating Spreadsheet Data via Accurate and Low-Effort Extraction

Zhe (Shirley) Chen and Michael Cafarella
University of Michigan, Ann Arbor

KDD 2014
This year's special theme: Data Science for Social Good
20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining
August 24-27, 2014 New York City

Introduction

❖Spreadsheets are a critical and widely-used data management tool: Microsoft estimates the number of worldwide Excel users at *more than 400 million*. Moreover, the *U.S. government* for many years published thousands of spreadsheets about economics, transportation, and etc.

❖If spreadsheet data can be easily converted to the *relational model*, many researchers --- in public policy, public health, economics and other areas --- could benefit from society's huge investment in relational integration tool.

❖However, we observe that a great amount of Web spreadsheets exhibiting a *hierarchical structure*.

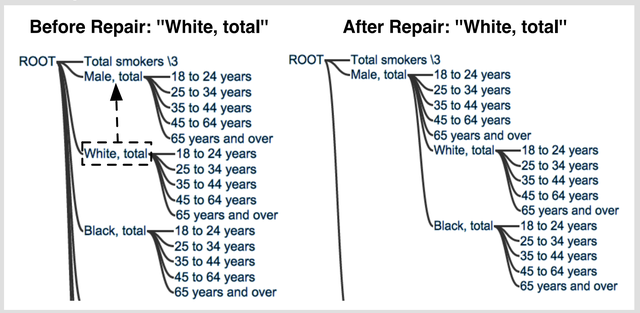
	Sex, age, and race	1990 \1	2000
19	**Total smokers \3**	25.5	23.2
20	Male, total	28.4	25.6
21	18 to 24 years	26.6	28.1
22	25 to 34 years	31.6	28.9
23	35 to 44 years	34.5	30.2
24	45 to 64 years	29.3	26.4
25	65 years and over	14.6	10.2
26	White, total	28.0	25.7
27	18 to 24 years	27.4	30.4
28	25 to 34 years	31.6	29.7
29	35 to 44 years	33.5	30.6
30	45 to 64 years	28.7	25.8
31	65 years and over	13.7	9.8
32	Black, total	32.5	26.2
33	18 to 24 years	21.3	20.9
34	25 to 34 years	33.8	23.2
35	35 to 44 years	42.0	30.7
36	45 to 64 years	36.7	32.2
37	65 years and over	21.5	14.2

1990	Male	White	45 to 64 years	28.7
1990	Male	White	65 years and over	13.7
2000	Male	White	45 to 64 years	25.8
2000	Male	Black	65 years and over	14.2

Interactive User Model

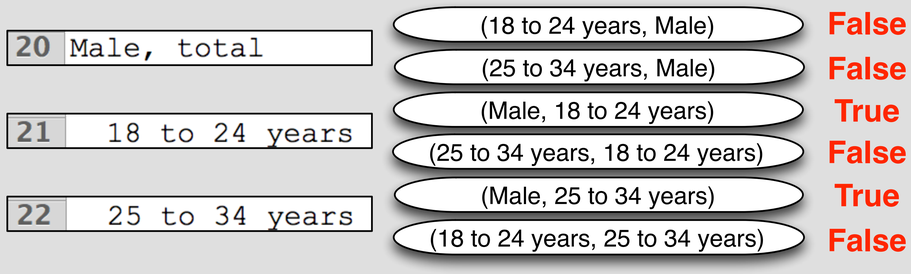
Our interactive user model is an iterative three-step proces:

1. **Initial Results**. The system presents the initial extraction results computed by the automatic extraction.
2. **Review and Repair.** A user can review the results on the interface and repair an annotation's parent from one to another via a dragging and dropping operation.
3. **Spread Repairs.** The system spreads the user's repair to many other similar mistakes.



An Undirected Graphical Model based Approach

❖**Encoding Hierarchy Extraction.** We create a Boolean variable x = (a_p, a_c) to represent a ParentChild pair candidate. Each variable x takes a label l from {true,false}, and x holds true if a_p is the parent of a_c.



❖**Correlating ParentChild Decisions.**

❖**Stylistic Affinity ---** Two ParentChild variables in one spreadsheet have identical visual style for parents and for children, are likely to share the same label.

E.g. The two nodes (Male, 18 to 24 years old) and (Male, White).

❖**Metadata Affinity ---** We can use metadata resource to find additional correlation between ParentChild variables both within and between spreadsheets.

E.g. The two nodes (White, 18 to 24 years old) and (Black, 25 to 34 years old).

❖**Adjacent Dependency ---** If we consider the ParentChild pairs of a single spreadsheet as a sequence, adjacent variables follows a transition pattern.

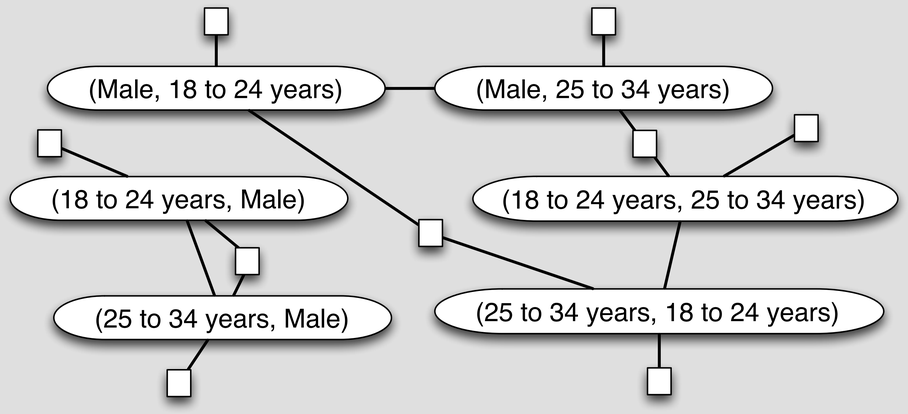
Aggregate Design --- Further constraints that reflect typical overall spreadsheet design and ensure that the resulting variable assignment yields a legal hierarchy.

❖**Orientation constraint**

❖**One-parent constraint**

Spreadsheet Hierarchy Extraction Graphical Model

❖**We encode the above observations as, node potentials, edge potentials, global potentials, and repair potentials.**



❖**Summary.** Let G be a graphical model that has a set of variables **x** = {x_1, …, x_n} where each x_i in **x** represents a ParentChild candidate in an annotation region and takes a label l_i from L = {true, false}. Let $\mathbf{l_r}$ be the set of repair-induced labels on variables $\mathbf{x_r}$. The joint distribution of G is:

$$P(\mathbf{l} \mid \mathbf{l_r}, \mathbf{x}) = \frac{1}{Z(\mathbf{w})} exp\Big(\sum_x \theta(x,l) + \sum_x \sum_{x'} \theta(x,l,x',l') + \sum_{k \in \{a,b\}} \phi_k(\mathbf{x},\mathbf{l}) + \sum_x \sum_{x_r \in \mathbf{x_r}} \varphi(x,l,x_r,l_r)\Big)$$

Experiments

❖**We use two spreadsheet corpora.**

❖SAUS --- The 2010 Statistical Abstract of the United States consists of 1369 spreadsheet files totaling 70MB.

❖WEB --- Our Web dataset consists of 410,554 Microsoft Excel files from 51252 distinct Internet domains, totaling 101GB.

❖**Automatic Extraction.**

❖**AutoBasic** – The baseline.

❖**AutoLR** --- Node potential only.

❖**AutoEdge** --- Node + edge potential.

❖**AutoGlobal** --- Node + global potential

❖**AutoFull** --- Node + edge + global potential.

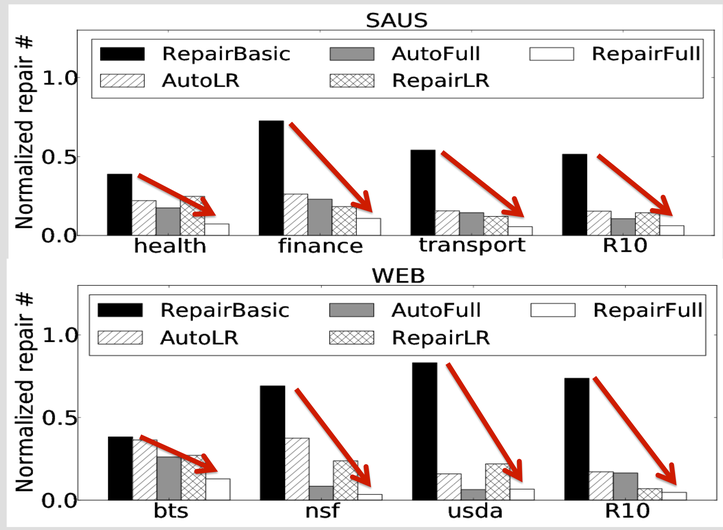
	Methods	F1
SAUS	AutoBasic	0.4641
	AutoLR	0.8751
	AutoEdge	0.8794
	AutoGlobal	0.8834
	AutoFull	0.8860
WEB	AutoBasic	0.4736
	AutoLR	0.7892
	AutoEdge	0.7973
	AutoGlobal	0.8122
	AutoFull	0.8327

❖**Interactive Repair.**

❖**RepairBasic** --- The baseline.

❖**RepairLR** --- Node + repair potential.

❖**RepairFull** --- Node + edge + global + repair potential.



References

[1] Zhe Chen, Mike Cafarella: Integrating Spreadsheet Data via Accurate and Low-Effort Extraction. KDD (2014).
[2] Zhe Chen, Mike Cafarella, Jun Chen, Daniel Prevo, Junfeng Zhuang: Senbazuru: A Prototype Spreadsheet Database Management System. PVLDB 6(12): 1202-1205 (2013).
[3] Zhe Chen, Michael J. Cafarella: Automatic web spreadsheet data extraction. SSW@VLDB 2013: 1.

Electrical Engineering and Computer Science
University of Michigan ■ College of Engineering

Contact: chenzhe@umich.edu