Senbazuru: A Prototype Spreadsheet Database Management System

Zhe Chen, Michael Cafarella, Jun Chen, Daniel Prevo, Junfeng Zhuang
Computer Science Department, University of Michigan

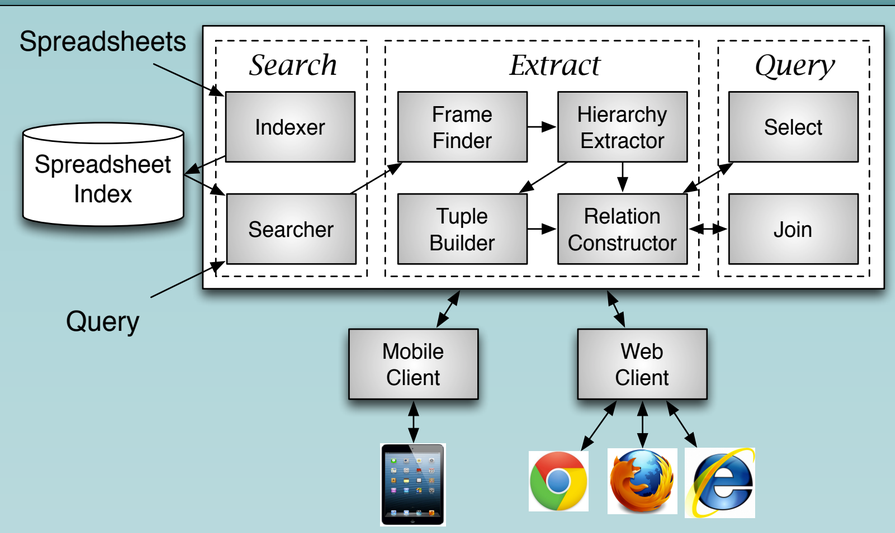
Abstract

Senbazuru is a prototype spreadsheet database management system (SSDBMS). It is able to extract relational information from spreadsheets, which opens up opportunities for integration among spreadsheets and with other relational sources.

We demonstrate that Senbazuru allows users to search for relevant spreadsheets in a large corpus, probabilistically constructs a relational version of the data, and offers several relational operations over the resulting extracted data (including joins to other spreadsheet data).

Our demonstration is available on two clients: a JavaScript website and a touch interface on the iPad.

Framework



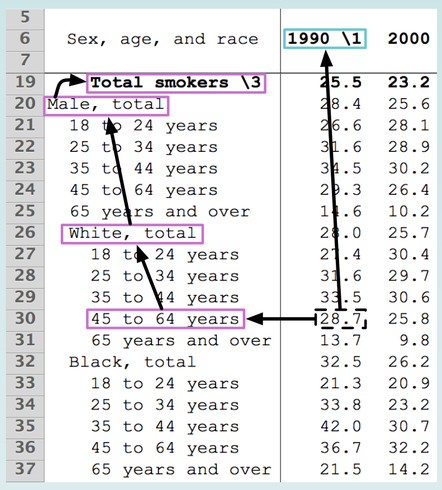
Senbazuru consists of three functional components we view as critical for a useful SSDBMS:

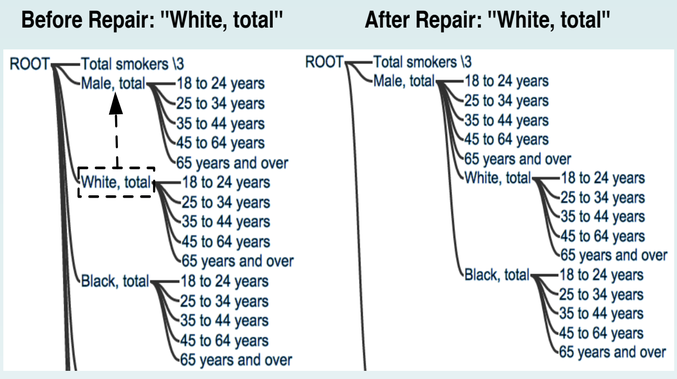
- **Search** -- For each spreadsheet in the corpus, the **indexer** extract text from each cell and uses Lucene to index the text. When a query arrives, the **searcher** uses the inverted index to find relevant datasets.
- **Extract** -- The extract component is composed of a background extraction pipeline that automatically obtains relational data from spreadsheets, and a repair interface that allows users to manually repair extraction errors.

Interface

Search -- Using a textual search-and-rank interface, the search component allows a user to quickly locate relevant datasets in a huge Web spreadsheet corpus.

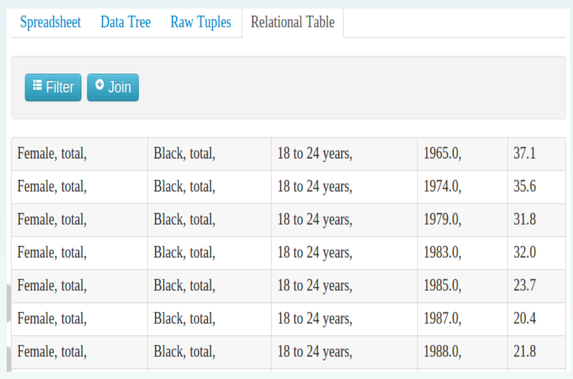
Extract -- Spreadsheets often exhibit a implicit hierarchical structure between attributes and values. The extract component automatically infer the implicit structure of spreadsheets based on a probabilistic model.





Repair -- Automatic extraction often emits errors, but our interface allows users to manually repair extraction errors effectively and efficiently.

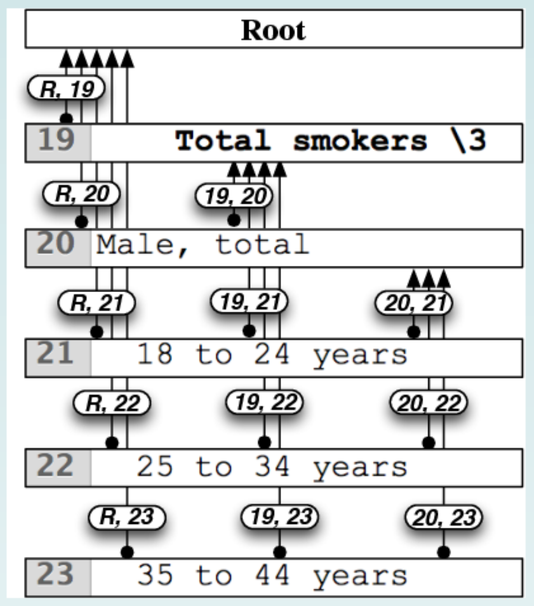
Query – Our interface supports relational query operations on the derived relational tables, such as filtering (i.e., selection).

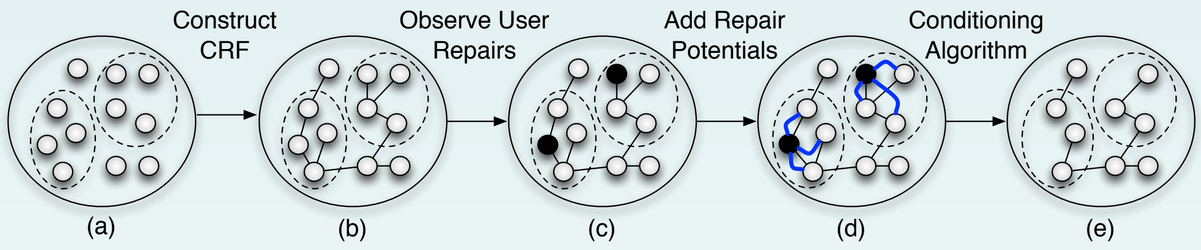


Integrate – Users are also able to integrate two spreadsheets based on the derived relational tables.

- **Frame Finder** – It identifies the attribute and value regions in a data frame spreadsheet.

- **Hierarchy Extractor** – It recovers the implicit attribute hierarchies for each attribute region.
  - Create random variables.
  - Establish the correlation between variables.
  - Incorporate user repairs.





- **Tuple Builder** – It generates a relational tuple for each value in the value region.

1990	Male	White	45 to 64 years	65 years and over	28.7

- **Relation Constructor** – It assembles the relational tuples into relational tables.

- **Query** -- The query component supports basic relational operators, especially **selection** and **join**, which the user can apply to spreadsheet-derived relations.

ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
UNIVERSITY of MICHIGAN ■ COLLEGE of ENGINEERING

Contact: chenzhe@umich.edu