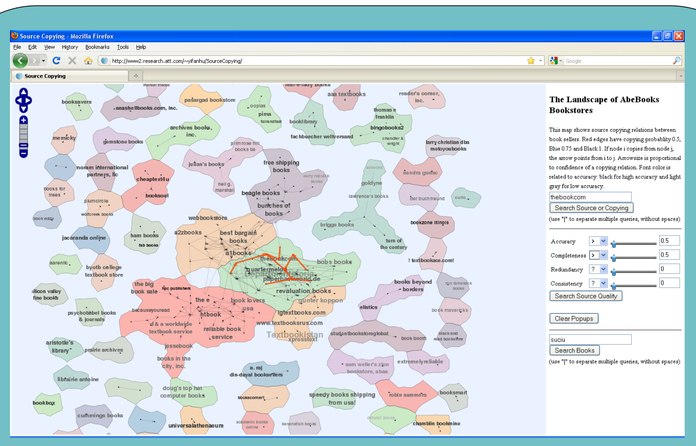
SOLOMON

Seeking the Truth Via Copying Detection



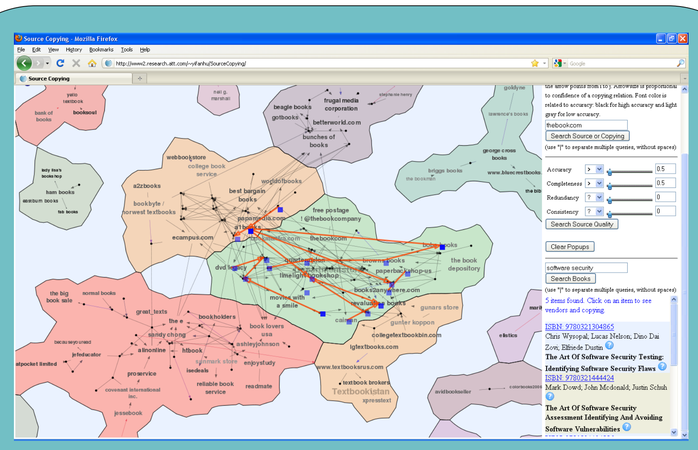
Getting Started: SOLOMON shows the data sources in a map: each "node" represents a data source; an edge "$S_1 \rightarrow S_2$" indicates that S_1 copies from S_2; each "country" represents a cluster of sources according to their copying relationships. SOLOMON supports source search (by name or by quality) and copying search. The red arrows highlight the results of searching (direct or transitive) copiers of "*TheBookCom*".

Motivation

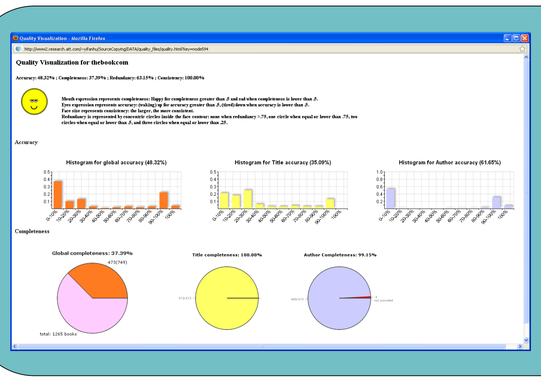
We live in the Information Era, with access to a huge amount of information from a variety of data sources. However, data sources are of different qualities, often providing conflicting, out-of-date and incomplete data. Data sources can also easily copy data from others, propagating erroneous data. Thus, identifying high quality information and sources is non-trivial.

We demonstrate SOLOMON, whose core is a module that detects copying between sources. We demonstrate that we can effectively detect copying, leverage the results in truth discovery, and provide a user-friendly interface to assist users in identifying sources that best suit their information needs.

Data set: Data extracted by searching computer science books from an online bookstore (source) aggregator, *AbeBooks,* in 2007 [4].

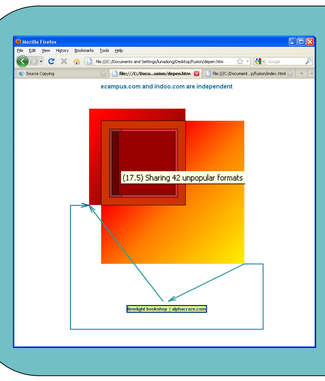


Data Search: SOLOMON supports keyword search on data items. The snapshot shows the results of searching "*software security*": the bottom right shows the returned book list, the blue spots show the providers of data for the book with ISBN *978032144424*, and the red arrows show the copying relationships between those data providers.



Quality Visualization:

SOLOMON visualizes various quality measures of a source using a Chernoff face. It also provides explanations for *completeness* of data using pie charts of provided items, and explanations for *accuracy* of data using histograms of the probabilities that each provided value is true.

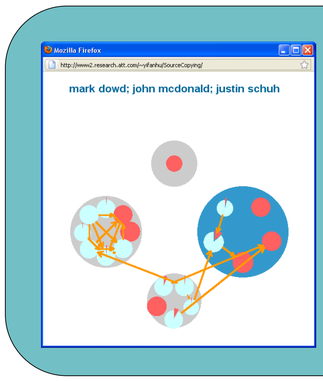


Explanation of (Non-)Copying:

SOLOMON explains copying-detection decisions.

The local decision (*indoo.com* copying from *ecampus.com*) is explained using the squares: the two squares each represents a data source; the overlapping area represents the overlapping data items; inside the overlapping area, the larger square represents shared values and the smaller square represents shared formats; vertical lines divide unpopular elements (strong evidence for copying) and popular elements (weak evidence); color of the squares encodes quality difference.

The global decision (two sources being independent) is explained by the transitive copying through sources *limelight bookshop* and *alphacraze.com*.



Explanation of Truth Discovery:

SOLOMON explains truth discovery decisions. In the figure, the blue circle represents the correct value and the grey circles represent incorrect values, where the size of the circle is proportional to the vote count. Each small circle inside the big circles represents a data provider, where the size of the circle is proportional to the accuracy. Each edge represents copying of the value between two providers and each red pie represents the final vote count from each source.

This example shows that SOLOMON ignores copied values and will not be biased by an incorrect value that is copied by many sources.

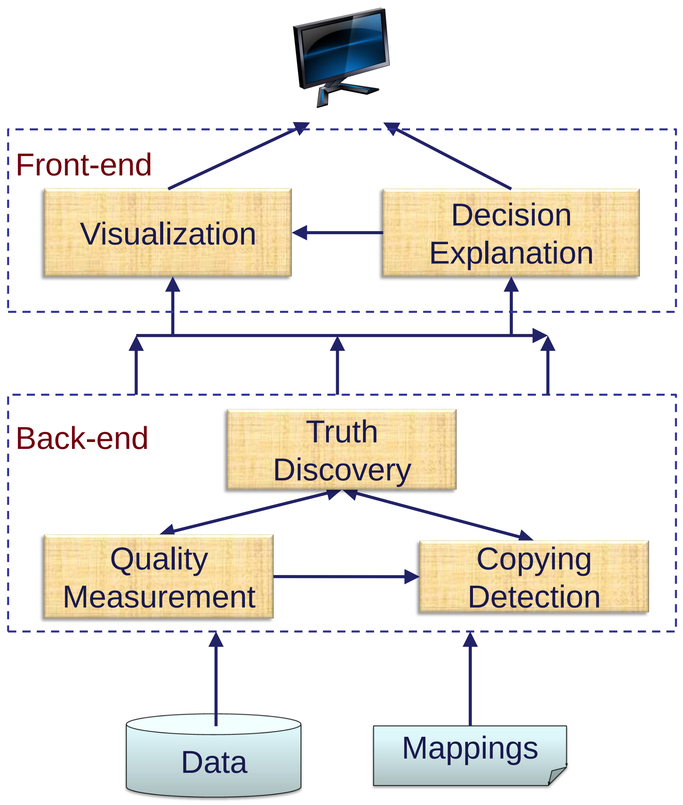
Back-End

The back-end of SOLOMON takes the data from various sources and the schema mappings as input, performs data fusion and infers quality measures of sources, copying relationships between sources, and true values for each data item.

Copying Detection: Copying detection proceeds in two steps [1,2]. The first step, *local detection,* discovers copying for each pair of sources independently of other sources. The second step, *global detection*, identifies co-copying and transitive copying.

Truth Discovery: When deciding the truth from conflicting values, SOLOMON not only considers which value the majority of sources vote for, but also ignores the copied values and gives higher weight to data from more accurate sources [1].

Quality Measurement: SOLOMON measures the quality of sources by *completeness* (the percentage of data that are provided), *accuracy* (correctness of the provided values), *consistency* (number of distinct values provided for each item), and *redundancy* (number of records provided for each object) [2].



Front-End

The front-end of SOLOMON provides a search and browsing interface to the user, generating visualizations and explanations on users' demand. The web-based front-end is written in JavaScript and thus completely portable.

Decision Explanation: A user often wonders not only "what" but also "why". SOLOMON provides explanation of various decisions, interpreting the underlying Bayesian analysis in a way that non-technical users can understand.

Visualization: SOLOMON provides an effective visualization to assist understanding of source quality and copying relationship. It applies the GMap techniques [3] and shows the sources in a map where closely related (by copying) sources are put close to each other. It also provides visualization for explanation of various decisions.

References

[1] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. *PVLDB,* 2009.

[2] X. L. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. Global detection of complex copying relationships between sources. *PVLDB,* 2010.

[3] E. Gansner, Y. Hu, and S. Kobourov. GMap: Visualizing graphs and clusters as map. In *IEEE Pacific Visualization Symposium,* 2010.

[4] X. Yin, J. Han, and P.S. Yu. Truth discovery with multiple conflicting information providers on the Web. *SIGKDD*, 2007.

Demo URL: http://www2.research.att.com/~yifanhu/SourceCopying/

Authors: Xin Luna Dong, Laure Berti-Equille, Yifan Hu, and Divesh Srivastava