Discovering the History of Data

Argyro Avgoustaki, Giorgos Flouris, Irini Fundulaki, Dimitris Plexousakis

{argiro,fgeo,fundul,dp}@ics.forth.gr

Institute of Computer Science – FORTH

Motivation

For some people the accuracy and the reliability of data are of high importance…

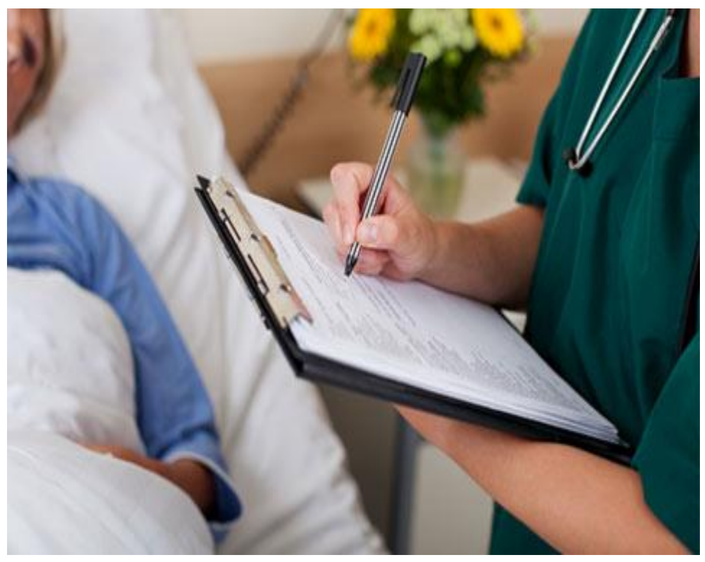
① The New York Times

A Mistaken News Report Hurts United

Unites Airlines Stock price after the fake bankruptcy

An erroneous report, saying United had filed for a second bankruptcy, sent its stock plummeting. By the end of the day, fingers were pointing in many directions to assign blame. United blamed an old Chicago Tribune article that, it said, was posted on the Web site of The South Florida Sun-Sentinel.

② ***Errors in patients' data entry***



Clinical data heavily rely on human entry and will do so for the foreseeable future. Medication errors happen all the time, an estimated one million each year, contributing to 7,000 deaths. On average there is one medication error every day for every inpatient.

Source: Forbes

③ ***Hoax Convinces Germany of Fake U.S. Suicide Bombing***

The hoax has transfixed this country. It prompted a 1,000-word tome on the website of Frankfurter Allgemeine Zeitung, Germany's most respected newspaper, and even a press conference denouncing the incident by the DPA – the German service responsible for first disseminating the news about the "attack."

Source: www.wired.com (2009)

④

On May, 1999, five US JDAM guided bombs hit the Chinese embassy in Belgrade. According to the U.S. government, the intention had been to bomb the nearby Yugoslav Federal Directorate for Supply and Procurement. CIA director testified that the CIA had identified the wrong coordinates for a Yugoslav military target on the same street

Source: Wikipedia, theguardian

Data Provenance

All these examples clearly indicate the need for a mechanism able to estimate the quality of data. Recording the *provenance* of data, i.e. their history, allows us to support applications such as:

- Data Quality
- Audit Trail
- Replication Recipes
- Copyrights
- Access Control

Our work

We introduced a new model for recording provenance in a fine-grained level. Thus, we are able to know from *where* and *how* each piece of information was generated. The proposed model is formulated using **algebraic expressions,** such as polynomials , comprised by **operators,** such as $\odot$, $\oplus$, and **identifiers,** such as c_i; identifiers represent a piece of data in the database system.

Assume the generated information c_0. Its provenance could be equal to:

- c_1

This is the **copy-paste** case, i.e., c_0 was generated by copying c_1

- $(c_5 \odot c_8) \odot c_2$

c_0 was resulted by **join** ($\odot$) information of c_5, c_8 and c_2.

- $c_1 \oplus c_2$

c_0 was resulted by **union**($\oplus$) the same information coming from different contributors.

Use Case Provenance Example

Assume the Use Case ③. In a provenance-aware world DPA could have assessed the quality of the incoming misleading information instead of trusting Wikipedia. For instance, we could identify that:

- Bluewater is not even a real city in California
- Berlin Boys is a non existent rap group.
- Wikipedia entries was fake
- Public safety California phone numbers were fake
- Local TV station was fake

References

1. A. Avgoustaki, G. Flouris, I. Fundulaki, and D. Plexousakis. Provenance Management for Evolving RDF datasets. In ESWC, 2016.
2. A. Avgoustaki. Provenance management for SPARQL updates. Master's thesis, University of Crete, 2014.
3. G. Flouris, I. Fundulaki, P. Pediaditis, Y. Theoharis, and V. Christophides. Coloring RDF triples to capture provenance. In ISWC, 2009.

This work is funded and supported by

ΕΛ.Ι.Δ.Ε.Κ. Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας

FORTH INSTITUTE OF COMPUTER SCIENCE