Modeling the Evolving World

Anja Gruenheid

Systems Group
Department of Computer Science
ETH Zurich

anja.gruenheid@inf.ethz.ch

ABSTRACT

All around us, the world is changing continuously, evolving every minute of the day. These changes are captured in event data that are partial snapshots of its evolution. The idea of story evolution is to combine these bits of information, reconstructing the real world and its evolution through a digital model of their relationships.

1. INTRODUCTION

Entity resolution has been in the focus of data integration research for a long period of time, [1]. Part of the entity resolution problem is what is commonly referred to as entity evolution which is the idea that entities change over time, [2]. This notion gives entity resolution a new level of complexity as it adds a temporal dimension to it. In this work, we propose that the notion of evolution does not hold for entities only but also for the relationships between entities and discuss how they can be modelled and integrated with traditional entity resolution. Specifically, the idea of modelling the world entails a new take on the entity-relationship model that is established by interpreting digitally recorded events.

2. MODELLING THE WORLD

The abstract notion of an entity in this work corresponds to real-life people, organizations, countries, etc. Relationships between these entities then describe entity interactions. Similar to entities, relationships may evolve over time. Moreover, relationships between entities have multiple facets. For example countries have political as well as economic relationships with each other. We call these varying facets of entity relationships *stories*. Take as example Fig. 1 which shows the story coverage of the plane crash in the Ukraine by the Huffington Post and the Wall Street Journal.

So far, stories have been collected unconnectedly in the form of events. Event data is extracted from newspaper articles, blogs, and other available online content using natural language processing techniques. Events are commonly stored in a tuple format, i.e., an event is identified by a source and a target entity and describes an interaction between them. In Fig. 1, an example event is <Separatists, Ukraine, Breaking News: Plane Crash>. Story identification is then the task of connecting events over time where connected events have

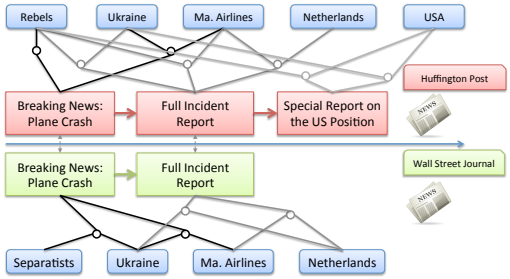

Figure 1: Evolving information as observed by two sources.

a continuous and consistent semantic relationship. Furthermore, when representing stories, we need to account for variations depending on who tells the story. We call the channels, through which events are reported *sources*. For example, each newspaper or blog can be viewed as a source. The abundance of sources raises several challenges for data cleaning and integration: Some sources provide biased information, they can provide additional or less information than other sources, or their availability may change over time [3].

To tackle the challenging task of modeling story evolution, we propose a two-step strategy: First, stories are identified and established within a data source. Consistent entity resolution as well as methods for story identification are required as part of this step. That is, within each source, data extraction methods establish the events specific to this source which are then evaluated in the context of previous events with the same or similar sets of associated entities and content. Second, stories are integrated across sources to provide complete information on a certain topic. For example in Fig. 1, only the Huffington Post reports on the US' position in the Ukraine conflict while the Wall Street Journal provides the information that the rebels are in fact separatists. Cross-source integration can be achieved by aligning stories across sources.

3. CONCLUSION

We introduced the problem of modelling the evolution of the world as a novel way of interpreting event data to determine entities and their relationships from observed data. We briefly discussed how stories can be extracted by first identifying stories within each source and then integrating the gained knowledge across sources to form a better representation of the world.

4. REFERENCES

[1] H. Garcia-Molina. Entity Resolution: Overview and Challenges. In *Conceptual Modeling - ER 2004*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2004.

[2] P. Li, X. L. Dong, A. Maurino, and D. Srivastava. Linking Temporal Records. *PVLDB*, 4(11):956–967, 2011.

[3] Nicholas Weller and Kenneth McCubbins. Raining on the Parade: Some Cautions Regarding the GDELT Dataset, 2014.

7th Biennial Conference on Innovative Data Systems Research (CIDR '15) January 4-7, 2015, Asilomar, California, USA.