A Time Machine for Information: Looking Back to Look Forward

Xin Luna Dong
Google Inc.
lunadong@google.com

Wang-Chiew Tan*
UC Santa Cruz
tan@cs.ucsc.edu

1. INTRODUCTION

> *"The longer you can look back, the farther you can look forward."* — Winston Churchill

With the abundant availability of information one can mine from the Web today, there is increasing interest to develop a complete understanding of the history of an entity (i.e., a person, a company, a music genre, a country, etc.) (see, for example, [7, 9, 10, 11]) and to depict trends over time [5, 12, 13]. This, however, remains a largely difficult and manual task despite more than a couple of decades of research in the areas of temporal databases and data integration.

The difficulty to create a comprehensive understanding of entities over time largely stems from the lack of (explicit) temporal data, and tools for interpreting such data even if they were available. Ideally, we would like to develop a time machine for information, where one can easily and incrementally ingest temporal data to form a more and more comprehensive understanding of entities over time, search and query facts for a particular time period, understand trending patterns over time, and perform analytics that would allow one to, for example, understand the prevalent "knowledge" in the previous decade. In this tutorial, we describe the techniques critical in building such a *time machine for information*, and discuss how far (or close) we are in achieving this goal.

The task of developing a time machine for information, where one can understand when a fact is true, for how long, and even when this was determined, is non-trivial. For one thing, the development of such a time machine would necessarily involve many of the challenges that occur in data integration [2, 4] and knowledge curation (see, for example, [1, 3, 8]), which are notoriously difficult tasks of their own. The data integration process include collecting information about different types of (heterogeneous) entities, transforming and cleaning the information, and curating facts regarding different aspects or properties of those entities consistently together. An additional challenge today is to perform these tasks at scale; that is, we need to collect information from a large number of data sources, where each source may contain lots of data, and the schemas of the sources may be diverse in their structure and quality. The ability to inter-operate amongst heterogeneous data sources with varying quality is thus a key ingredient to the successful development of this time machine.

Another key ingredient to the successful development of this time machine is to make every step of the data integration process *time-aware*. In other words, we need to understand the valid time period for each piece of fact and even when the validity was made known. To achieve this goal, one would inevitably require text extraction rules or techniques to extract structured temporal data from unstructured and semi-structured data sources. Furthermore, the extracted temporal data has to be mapped and transformed into a desired format before temporal entity resolution is applied. And finally, information about the extracted entities is integrated and the conflicting information is resolved to arrive at an integrated archive. This process may repeat as new datasets are discovered or when new versions of the same datasets are available to further enrich the information time machine.

In this tutorial we first introduce and motivate the need to develop a time machine for information with our examples and use cases. We will then survey and present existing work on three components (extraction, linking, and fusion) that are central to the development of any time machine for information before we conclude with our thoughts of what are some of the interesting open research problems. While one goal of this tutorial is to disseminate the above described material, a parallel goal is to motivate the audience to pursue research in the direction of managing and integrating temporal data through our tutorial. Ultimately, we hope to bring the research community and the industry one step closer to realizing the goal of building a time machine for information that will record and preserve history accurately, and to help people "look back" and so as to "look forward".

2. OUTLINE OF OUR TUTORIAL

2.1 Introduction and motivation

Our tutorial contains a number of examples to illustrate that search and data exploration are limited by the current data and knowledge management techniques. Some examples include answers to the query "Google CEO in 2015" (current Google search returns a speech by the "ex-CEO"), and answers to the query "Google's CEO before Larry Page" (current Google search returns articles about Larry Page), the difficulty in understanding the lengthy articles for "history of Google", and the difficulty to figure out a complete picture for Google employee growth.

Another compelling example to motivate the need to understand *when* a fact is true comes from reports that are filed with the U.S. Securities and Exchange Commission (SEC) [6] at different times.

*Tan is partially supported by NSF Grant IIS-1450560.

Proceedings of the VLDB Endowment, Vol. 8, No. 12
Copyright 2015 VLDB Endowment 2150-8097/15/08.

Companies are required, by federal regulations, to file reports periodically to SEC to disclose the stock holdings of its executives. There are now millions of electronic filings in EDGAR and the number of such filings is increasing over time. Given the millions of SEC reports, how can one find out the stock holdings of an executive during a certain period of time? Were *Ann* and *Bob* affiliated with the same company during a certain time period? Or perhaps more interestingly, did *Ann* purchase a significant number of shares of *Company X* before it was announced that *Company X* would be bought by *Company Y*?

We then present several efforts that facilitate the exploration of temporal or historical data by integrating temporal data sources.

2.2 Connecting back to Temporal Databases

Through the examples, we show that building a time machine for information necessarily calls for two additional types of information to be curated. First, each fact must be associated with the time period(s) over which it is valid. For example, the fact that Cormen, Leisersen, and Rivest were authors of the first edition of the "Introduction to Algorithms" book is true since 1990 and this fact will never change, even though there is a second edition with the same authors, and a third edition with an additional author Stein. On the other hand, some facts are true only for a finite time period or a single time point. For example, the CEO of Google Inc. was Eric Schmidt from 2001 to 2011, neither before nor after.

Second, it is also important to identify how perspectives change over time, so history can be accurately captured. As an example, while the earth has always been round and revolving around the sun, there was a time when people believed that the earth is the center of the universe and that the earth is flat.

We very briefly review temporal databases, state how these two types of temporal information may be effectively expressed using valid time and transaction time of bi-temporal databases,and discuss some limitations thereof.

2.3 Extracting, Linking, Fusing Temporal Data

Extraction Extracting temporal data is difficult for three main reasons. First, we need to recognize the occurrences of temporality in data; often times there is no explicit time stamp such as "May 4th, 2014", but implicit ones such as "one year later" and "before she joined the band". Second, one also needs to understand the relationship between the fact at hand and the timestamp, which are often separated by a long context. Finally, in addition to explicitly mentioned time stamps, one can also mine the time of the events indirectly. For example, from the edit history which is available in data sources such as Wikipedia and some websites.

We describe three bodies of works for extraction that focuses on (1) natural language processing and free text, (2) (semi-)structured data, such as the bibliography-style Wikipedia pages and from personal resumes, (3) and edit histories.

Linking Since temporal data often span a long period of time, there are two fundamental challenges that arise in temporal record linkage but do not exist in linking snapshot data. First, when linking two records, one has to consider the evolution of the status and attribute values of an entity over time; that is, the same person may change affiliation and the same company may change the location of its headquarter. Second, as we have a long history and presumably many more entities, it is likely that one will observe different entities with the same attribute values; for example, the possibility of finding different people with the same name and affiliation is much higher considering the whole history of the affiliation than considering a particular time point.

We will present recent work on linking temporal records.

Fusion There are two fundamental challenges in fusing temporal data. First, the time aspect associated with data is often open to interpretation. For example, a colloquium announcement saying "Professor Smith from Stanford gives a talk on March 22, 2015" asserts that Smith was in Stanford in March 2015, but does not state the time period when she worked in Stanford. Second, as in traditional data integration, inconsistencies may arise with respect to temporal constraints when data from multiple sources are combined together. The data not only can be imprecise, but also can be out-of-date. Continuing with our earlier example, a Berkeley project page in the same time frame listing the same *Smith* as a faculty member is likely to contain out-of-date information. Which data are correct depend not only on the accuracy of the source but also freshness of the information from the source.

We will overview past research on conflict resolution.

2.4 Conclusion and open research problems

We conclude our tutorial by stating a number of open problems we need to solve to move towards the goal of building a time machine for information. The open problems include the need to leverage existing work on temporal databases to model and store the data, addressing the additional challenges we face for extracting, linking, and fusing data, building history not only for individual entities such as a person, but also for collective entities such as a country and a war, distinguishing fact evolution and perspective evolution, presenting the rich history to people in the most understandable way, and so on.

We hope our tutorial can inspire and advance research in this important area, and bring us one step closer to realizing the dream of building an information time machine.

3. REFERENCES

[1] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, pages 1247–1250, 2008.

[2] A. Doan, A. Halevy, and Z. Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012.

[3] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *SIGKDD*, 2014.

[4] X. L. Dong and D. Srivastava. *Big Data Integration*. Morgan & Claypool, 2015.

[5] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. *ACM TWEB*, 1(2):7, 2007.

[6] The EDGAR Public Dissemination Service. `http://www.sec.gov/edgar.shtml`.

[7] D. Graus, M.-H. Peetz, D. Odijk, O. de Rooij, and M. de Rijke. yourhistory–semantic linking for a personalized timeline of historic events. *Workshop: LinkedUp Challenge at Open Knowledge Conference (OKCon) 2013*, 2013.

[8] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *AIJ*, 2012.

[9] J. Li and C. Cardie. Timeline generation: tracking individuals on twitter. In *WWW*, pages 643–652, 2014.

[10] A. Mazeika, T. Tylenda, and G. Weikum. Entity timelines: Visual analytics and named entity evolution. In *ACM CIKM*, pages 2585–2588. ACM, 2011.

[11] M. Roth and W.-C. Tan. Data integration and data exchange: It's really about time. In *CIDR*, 2013.

[12] Y. Wang, M. Zhu, L. Qu, M. Spaniol, and G. Weikum. Timely yago: harvesting, querying, and visualizing temporal knowledge from wikipedia. In *EDBT*, pages 697–700. ACM, 2010.

[13] G. Weikum, N. Ntarmos, M. Spaniol, P. Triantafillou, A. A. Benczúr, S. Kirkpatrick, P. Rigaux, and M. Williamson. Longitudinal analytics on web archive data: It's about time! In *CIDR*, pages 199–202, 2011.