CLTR: Collectively Linking Temporal Records Across Heterogeneous Sources

Yanyan Zou(✉) and Kasun S. Perera

Singapore University of Technology and Design, Singapore, Singapore
yanyan_zou@mymail.sutd.edu.sg, baruhupolage@sutd.edu.sg

Abstract. A huge volume of data on the Web are continually made available, which provides users rich amount of information to learn more about entities. In addition to attribute values of entities, there is often additional relational information, such as friendship on social networks, coauthorship of papers. However, to understand how these facts across heterogeneous data sources are related is challenging for users due to entity evolution over time. In this paper, we propose a novel system to help users find how records are temporally related and understand how entity profiles evolve over time. Our system is able to Collectively Link Temporal Records (CLTR) by taking advantage of evidence from both attribute and relational information on multiple sources. We demonstrate how CLTR allows users to explore time-varying history of targeted entities and visualizes multi-type relations among entities.

Keywords: Record linkage · Temporal data · Collective clustering · Heterogeneous web sources

1 Introduction

Heterogeneous Web sources provide abundant information to describe entities from different aspects over a long period of time. In addition to attribute values of entities, there is often additional relational information on the Web, such as friendship on social networks, coauthorship of papers. Understanding how facts across heterogeneous data sources are temporally related is paramount and inevitable to many applications. It also poses new challenges due to evolving information of entities where attribute values may vary over time (e.g., location, age and organization). In order to deal with time-varying property of temporal records, Li et.al. first presented a time-decay model [1] which learns the probability that an entity will change its attribute values within a time period. CHRONOS [4], the closest to our work, is implemented based on this technique. It collects data only from bibliography domain. This work was then extended by [2], where a mutation model was proposed to capture how likely an entity will return to its previous values in a given time period. Another framework, called MAROON [3], was designed to observe temporal patterns of attribute value

S. Candan et al. (Eds.): DASFAA 2017, Part II, LNCS 10178, pp. 668–671, 2017.
DOI: 10.1007/978-3-319-55699-4_43

transitions. None of the models emphasizes relational information among entities, which actually provides us additional evidence to address temporal linkage across heterogeneous Web sources.

In this paper, we present a novel framework, CLTR, to help users understand how facts are temporally related across sources. The major contributions of CLTR can be summarized as follows:

1. CLTR collects data from multiple Web sources to enrich entity profiles.
2. The system combines evidence from both attribute and relational information to reconcile entities across heterogeneous web sources. It employs collective clustering mechanism. The algorithm fully utilizes relations to jointly reconcile entities which co-occur on the Web, rather than independently.
3. It enables users to search targeted entities by different domain keywords, such as name and organization. The tool presents entity history by timelines and also visualizes entity relation graph to help users further understand how entities are temporally related across sources over time.

2 Methodology

The architecture of our CLTR system is illustrated in Fig. 1. It takes as input data crawled from heterogeneous web sources and visualizes history profiles and relations of entities. The system consists of three key components: input and data preprocessing, temporal record linkage and interactive interface.

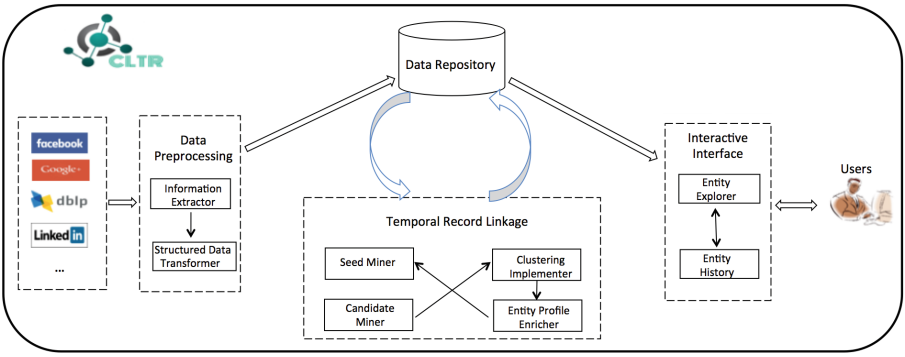


Fig. 1. Architecture of CLTR system

Data Preprocessing takes as input data provided by web sources, e.g., Facebook, Linkedin, DBLP, and personal homepages. This component transforms facts about entities into records associated with timestamps, including name, title, organization, interests, social connections and so on. It stores records in the data repository.

Temporal Record Linkage implements a collective clustering algorithm to determine related entities jointly. Initially, it discovers reconciled entities as seeds via linked web sources. Starting with seed entities, the component applies relations to further reconcile more entities which are related to them. It searches for potentially mergeable pairs from related records and pushes them into the clustering algorithm to determine whether they should be merged. Once new merging decisions are made, profiles of the corresponding entities are further enriched. On the other hand, these newly-merged records are considered as new seeds to enlarge seed entity set and to discover more candidates. The clustering results are stored in the data repository.

Relational information is beneficial to propagate evidence among records. It can also contribute to making merging decisions. To combine attribute information with relational evidence, we define a new metric:

$$Match(r_1, r_2) = \alpha\Phi(r_1, r_2) + (1 - \alpha)\Psi(N(r_1), N(r_2)) \quad (1)$$

where r_1, r_2 represent two records, $\Phi(r_1, r_2)$ denotes temporal similarity of attribute values, $\Psi(N(r_1), N(r_2))$ captures evidence from two records' relations, $N(r)$ denotes the neighbor of record r via relations, α is a balancing factor to assign weights for temporal and relational components.

Interactive Interface offers users the explorer to search by different attribute keywords, such as name, organization, and the timelines to trace the complete history for each entity.

3 Demonstration Scenario

Using a running example, we demonstrate how users can interact with our system to search for a specific entity.

Consider a user who would like to find a person named "Meng Jiang". He selects keyword domain and searches by name "meng jiang". Figure 2 depicts

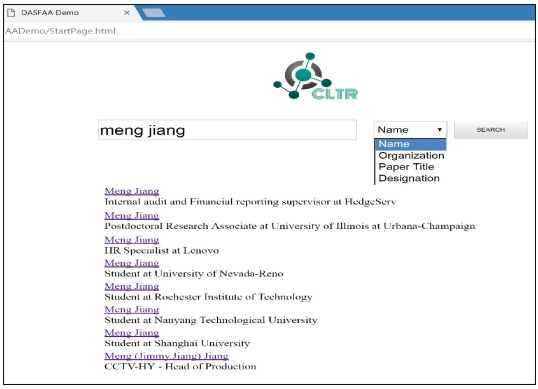
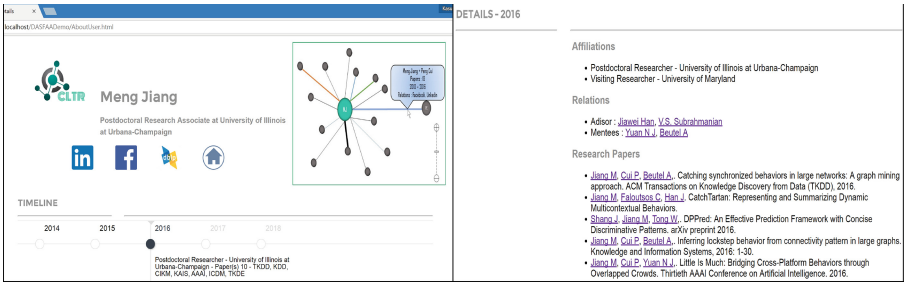


Fig. 2. Entity explorer interface



(a) Entiy timeline and relation graph (b) Entity's details of the selected time point

Fig. 3. Detailed entity's history and relations depiction

possible search results for this. Each result is shown with latest profile summary (i.e. title and affiliation) of fuzzy matched "meng jiang" in our datasets. The user can select one of them to trace more details.

Suppose the user has selected the second result, then the system switches to profiling page of this entity. As illustrated in Fig. 3a, it shows various aspects of this person, such as his most recent affiliation, title, linked websites according to data sources. The interactive timeline summarizes profiles of the entity over years, e.g., in 2016, "Meng Jiang" is a postdoctoral researcher at "University of Illionois at Urbana-Champaign", and has published 10 papers in 7 conferences. It is accessible for users to scroll over the timeline to switch time points. If the user wants to know more details in 2016, he can click the year node, as shown in Fig. 3b, "Meng Jiang" visited "University of Maryland" this year. Related entities to the selected one are shown on an interactive graph, depicted next to entity summary. For example, we show the relation graph of "Meng Jiang", related by his coworkers, social friends, etc. The thickness of edges is proportional to the strength of the relations. The color of edges represents relations from different sources (e.g., Facebook, DBLP, etc.). Clicking on each edge, user can observe more information about the entity relationship.

References

1. Li, P., Dong, X.L., Maurino, A., Srivastava, D.: Linking temporal records. Proc. VLDB Endow. **4**(11), 956–967 (2011)
2. Chiang, Y.H., Doan, A.H., Naughton, J.F.: Modeling entity evolution for temporal record matching. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, pp. 1175–1186. ACM (2014)
3. Li, F., Lee, M.L., Hsu, W., Tan, W.: Linking temporal records for profiling entities. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 593–605. ACM (2015)
4. Li, P., Tziviskou, C., Wang, H., Dong, X.L., Liu, X., Maurino, A., Srivastava, D.: Chronos: facilitating history discovery by linking temporal records. Proc. VLDB Endow. **5**(12), 2006–2009 (2015)