Information Systems 65 (2017) 124–136

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

Multi-source uncertain entity resolution: Transforming holocaust victim reports into people

Tomer Sagi[a,*], Avigdor Gal[b], Omer Barkol[a], Ruth Bergman[a], Alexander Avram[c]

[a] Hewlett Packard Labs, Guttwirt Industrial Park, Technion City, Haifa, Israel
[b] Technion - Israel Institute of Technology, Technion City, Haifa, Israel
[c] Yad Vashem, Jerusalem, Israel

ARTICLE INFO

Keywords:
Uncertain entity resolution
Blocking
Holocaust

ABSTRACT

In this work we present a multi-source uncertain entity resolution model and show its implementation in a use case of Yad Vashem, the central repository of Holocaust-era information. The Yad Vashem dataset is unique with respect to classic entity resolution, by virtue of being both massively multi-source and by requiring multi-level entity resolution. With today's abundance of information sources, this project motivates the use of multi-source resolution on a big-data scale. We instantiate the proposed model using the MFIBlocks entity resolution algorithm and a machine learning approach, based upon decision trees to transform soft clusters into ranked clustering of records, representing possible entities. An extensive empirical evaluation demonstrates the unique properties of this dataset that make it a good candidate for multi-source entity resolution. We conclude with proposing avenues for future research in this realm.

1. Introduction

Cultural heritage institutes are tasked with recording, researching, and preserving a culture, often after a severe catastrophe causing a heightened sense of urgency in preserving the records of this culture. These organizations have collected, over the years, troves of analog artifacts, including films, audio recordings, documents, and pictures. Digitization of these artifacts and extraction of metadata and texts, either manually or via OCR techniques, has created a deluge of raw data through which researchers can sift, attempting to create coherent narratives of a culture now extinct. Recent attempts such as the EHRI project [5] create infrastructure that enables researchers around the globe to access disparate sources of information using a unified interface with underlying common semantics. However, with the growing amounts of data, data integration problems arise. A first step towards recreating the story of a specific person, community, or place is the unification of all information pertaining to these entities, overcoming different source schemas, languages, political, and historical idiosyncrasies.

As an example, we bring the story of Guido and Massimo Foa. Fig. 1 shows a picture titled “Guido and Massimo Foa, Cuorgnè, 1944”. From the picture we can deduce that there were once a father and son named Guido and Massimo (who is who?) and in 1944 they resided in Cuorgnè, Italy. The Yad Vashem Names Project,[1] has been collecting Holocaust victim reports since 1953. Among these are three reports, whose extracted data is presented in Table 1.

Yad Vashem also commemorates non-Jewish individuals who risked their lives to save Jewish people during the Holocaust. One of those commemorated is Clotilde Boggio, who hid a child named Massimo from the Nazis in a village called Cuorgnè from 1944 to 1945. Taking into account the information in these four sources, a graph such as the one in Fig. 2 can be established. During the construction process, data integration decisions need to be taken, *e.g., do all of the rows in* Table 1 *refer to the same person?* Furthermore, extraction of the three records presented in Table 1 is not trivial. A simple query selecting those records whose first name is Guido and last name is Foa would have missed the third record, which nonetheless contains valuable information.

Weaving information to form narratives, stories told as a sequence of events, has traditionally been a manual process, performed by expert historians. For example, Massimo Foa, Guido's son, grew up to be a historian and wrote a book describing his parents' story, which enables the validation of the knowledge graph (Fig. 2). The challenge we tackle in this work is to create a robust automatic procedure to identify and collect all information pertaining to a single entity from over 500,000 sources in the Yad Vashem database, as a stepping stone towards

[*] Corresponding author.
E-mail address: ts.tomersagi@gmail.com (T. Sagi).

[1] http://www.yadvashem.org/yv/en/remembrance/names/ retrieved June, 2015

http://dx.doi.org/10.1016/j.is.2016.12.003
Received 17 December 2016; Accepted 19 December 2016
Available online 23 December 2016

Fig. 1. Guido and Massimo Foa, Cuorgnè, 1944.

Table 1
Three victim reports from the Yad Vashem Names Project DB.

BookID	First	Last	Gender	DOB	Birth	Permanent	Death	Spouse	Mother	Father
1016196	Guido	Foa	Male	02/08/36	Torino Italy	Torino Italy			Estela	Italo
1059654	Guido	Foa	Male	18/11/20	Torino Italy	Torino Italy	Auschwitz	Helena	Olga	Donato
1028769	Guido	Foy	Male	18/11/20	Turin Italy	Canischio Italy			Olga	Donato

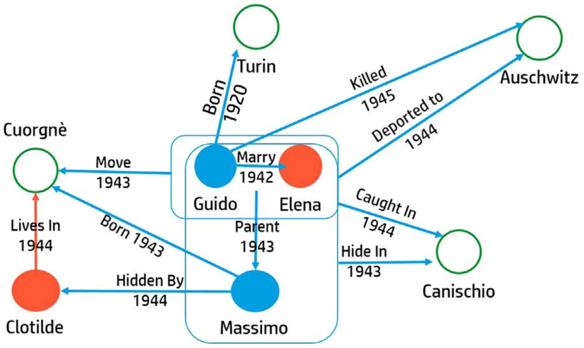

Fig. 2. Knowledge graph of Guido Foa.

automatically creating narratives for each entity in the database.

This data integration challenge can be positioned within the scope of the research area of entity resolution (also known as entity matching, record linkage, and deduplication) [29,11,19,22,9]. Entity resolution (ER) is at the heart of the data integration problem. The task entails creating a single entity from a collection of data records, each revealing some aspects of the entity, with no common identifier to rely upon. Examples include identifying accounts belonging to the same customer in different operational systems, merging customer accounts following a bank merger, *etc*.

Creating narratives from a set of facts poses a new challenge, which is non-characteristic of ER applications. Many ER applications require a single crisp answer as the outcome of the process, while here the outcome is a ranked list of possible narratives, which depends on the created ER clusters. Only in rare cases (such as the example given above) one would be lucky enough to find a single narrative that dominates the others. In most cases, we are faced with subjective details of events and based on the context may choose one narrative over another. To deal with the requirements as were set forward by the Yad Vashem application, we introduce a model for uncertain entity resolution, an ER process in which a tuple may be simultaneously associated with multiple entities. With uncertain entity resolution, entities are disambiguated only at query time, depending on the query at hand.

To support the flexibility uncertain ER requires, we use an entity blocking algorithm, MFIBlocks, to create soft blocks (clusters) and apply a machine learning method using decision trees to transform blocks into ranked associations of records to form entities. We also show how different queries affect the ER outcome differently.

This paper tells the story of cultural heritage institutions, such as Yad Vashem, and their effort to collect large amounts of information about the past. However, the multi-source entity resolution task may be relevant to any application that uses multiple disparate sources of information pertaining to the same people and events. While the described project was motivated by a desire to piece together stories from a lost culture, we believe its implications may benefit modern applications looking to automatically construct coherent narratives from a multitude of sources.

The work extends the work that was presented at SIGMOD'2016 [26] in two main directions. First, we provide a model for uncertain entity resolution and show its instantiation using a specific set of algorithms. Second, we significantly extend our empirical evaluation to show the suitability of the proposed solution for the uncertain entity resolution task.

The rest of the paper is structured as follows. Background on the Yad Vashem Names Project is given in Section 2. We then present a model for uncertain entity resolution (Section 3). The algorithmic solution we propose for uncertain entity resolution is provided in Section 4 followed by details of a concrete architecture for Yad Vashem in Section 5. We report on an extensive empirical evaluation of the model (Section 6) and discuss the implications of this work and avenues for future research in Section 7.

2. The Yad Vashem names project

In 1953 Yad Vashem was established as both a research institute and a memorial. One of its major tasks, beginning in 1954, was the registration of Holocaust victims' names on “Pages of Testimony” containing the names and biographic details of individual victims.

A national campaign of collecting Pages of Testimony in Israel between 1955-1957 resulted in 800,000 names registered by family members and friends. Collection efforts continued: during the 1980s the average number of incoming Pages of Testimony was 14,000-15,000 a year. Following the fall of the Iron Curtain and through the 1990s, the yearly average has risen to 30,000, largely in Russian. Pages of Testimony are preserved in the Hall of Names.

In September 1991 began the extraction of names and biographic data from the Pages of Testimony. This digitization project would later extend to all name resources in Yad Vashem, including the Archives and the Library. By the end of 1998, the Hall of Names digitized 470,000 Pages of Testimony. In addition, 500,000 names from major deportation lists were extracted through OCR.

In spring 1999, in agreement with the International Commission of Eminent Persons dealing with Swiss dormant bank accounts, Yad Vashem, Tadiran Systems Ltd., and Manpower Israel processed the remaining 1.1 million Pages of Testimony and scanned the entire collection. In parallel, in April 1999 Yad Vashem led a renewed campaign of collecting Pages of Testimony resulting in 400,000 Pages of Testimony and 50,000 photographs by the end of 2000.

The data was gathered into a database supported by a cataloging and retrieval system. Efforts have been made to streamline and standardize the data and to create advanced retrieval tools. The Central Database of Holocaust Victims' Names was launched on the Internet in November 2004 and is available in Hebrew, English,