Using the semantic web for author disambiguation - are we there yet?

Cornelia Hedeler[1], Bijan Parsia[1], and Brigitte Mathiak[2]

[1] School of Computer Science, The University of Manchester, Oxford Road, M13 9PL Manchester, UK,
{chedeler,bijan.parsia}@manchester.ac.uk
[2] GESIS - Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany
brigitte.mathiak@gesis.org

Abstract. The quality, and therefore, the usability and reliability of data in digital libraries depends on author disambiguation, i.e., the correct assignment of publications to a particular person. Author disambiguation aims to resolve name ambiguity, i.e., synonyms (the same author publishing under different names), and polysemes (different authors with the same name), and assign publications to the correct person. However, author disambiguation is difficult given that the information available in digital libraries is sparse and, when integrated from multiple data sources, contain inconsistencies in representation, e.g., of person names, or venue titles. Here we analyse and evaluate the usability of person-centred reference data available as linked data to complement the information present in digital libraries and aid author disambiguation.

1 Introduction

Users of digital libraries are not only interested in literature related to a particular topic or research field of interest, but more frequently also in literature written by a particular author [2]. However, as digital libraries tend to integrate information from various sources, they suffer from inconsistencies in representation of, e.g., author names or venue titles, despite best efforts to maintain a high data quality. For the actual disambiguation process, a wide variety of additional metadata are used, e.g., journal or conference names, author affiliations, co-author networks, and keywords or topics [1, 6]. However, in some digital libraries the available metadata can be quite sparse, providing insufficient amount and detail of information to disambiguate authors efficiently.

To complement the sometimes sparse bibliographic information a number of approaches surveyed in [1] utilise information available elsewhere, e.g., using web searches, and most of the approaches proposed are evaluated utilising gold standard datasets of high quality, such as Google Scholar author profiles. However, to the best of the authors' knowledge, these high quality data sets have so far not been used as part of the disambiguation process itself. Here we analyse person-centred reference data available on the semantic web and evaluate whether it contains sufficient detail and content to provide additional information and aid author disambiguation.

2 Data sets

2.1 Digital library data sets

In contrast to the wealth of metadata available in some digital libraries, the records in the two digital library data sets used here only offer limited metadata.
DBLP Most publication records in the DBLP Computer Science Bibliography [4] consist only of author names, publication titles, and venue information, such as names of conferences and journals. In addition to the publication records, DBLP also contains person records, which are created as result of ongoing efforts for author disambiguation [5].
Sowiport The portal Sowiport(http://sowiport.gesis.org) is provided by GESIS and contains publication records relevant to the social sciences. Here we only focus on a subset of just over 500,000 literature entries in Sowiport from three data sources (SOFIS, SOLIS, SSOAR) within GESIS, that have been annotated with keywords from TheSoz, a German thesaurus for the Social Sciences.

So far, no author disambiguation has taken place in these records, and inconsistencies in particular in author names make it hard for users to find all publications by a particular author. An analysis of the search logs has shown that the authors most frequently searched for are those with large numbers of publications, who tend to have entries in DBpedia and GND, motivating the use of the reference data sources introduced below.

2.2 Person-centred reference data

GND authority file and GND publication information. As the literature in Sowiport, in particular the subset used here, is heavily biased towards German literature, we use the Integrated Authority File (GND) of the German-speaking countries and the bibliographic data offered as part of the linked data service by the German National Library (http://www.dnb.de/EN/lds). Amongst other information, which also includes keywords, the GND file contains differentiated person records, which refer to a real person, and are used here.
DBpedia [3] is available for download(http://wiki.dbpedia.org/Downloads39) and comes in various data sets containing different kinds of data, amongst them 'Persondata', with information about people, such as their date and place of birth and death. As the persondata subset itself does not contain much additional detail, other data sets are required to obtain information useful for author disambiguation. The data is available either as raw infobox data or cleaned mapping-based data, which we use here.

3 Approach for author disambiguation

Our approach for author disambiguation can be seen as preliminary, as the main focus of this work was to evaluate whether there is sufficient information available in such reference data sets to make this a viable approach. It uses a domain specific heuristic as similarity function, and the reference data sets introduced above as additional (web information) evidence. To limit the number of records that need to be compared in detail, we use an index on the author/person names

Table 1. left: Number of person records in GND with selected professions; right: Number of instances in persondata in DBpedia for selected classes (y = yago).

Profession	# person
author/ female author	8,319 / 6,301
lecturer / female lecturer	7,931 / 1,204
research associated / female	1,117 / 759
physicist / female physicist	11,595 / 1,280
mathematician / female	7,561 / 908
computer scientist / female	5,443 / 589
sociologists / female sociologist	3,298 / 1,590
social scientist / female	998 / 546

Class	English	German
#foaf:person	1,055,682	479,201
#dbpedia-owl:person	652,031	215,585
#y:person	844,562	0
#dbpedia-owl:scientist	15,399	0
#y:Scientist110560637	44,033	0
#y:ComputerScientist109951070	1,667	0
#y:Mathematician110301261	4,994	0
#y:Physicist110428004	6,020	0
#y:SocialScientist110619642	9,083	0
#y:Philosopher110423589	6,116	0
#dbpedia-owl:Philosopher	1,276	0

Table 2. Number of person instances in DBpedia with selected properties of relevance.

Property	English	German
Author names		
foaf:Name	1,055,682	479,201
rdfs:label	1,055,682	479,201
dbpedia-owl:birthName	44,977	285
dbpedia-owl:pseudonym	1,865	0
Author affiliation		
dbpedia-owl:almaMater	42,318	0
dbpedia-owl:employer	3,232	0
dbpedia-owl:school	1,974	0
dbpedia-owl:university	1,073	0
dbpedia-owl:institution	923	0
dbpedia-owl:college	13,510	1,829

Property	English	German
Co-authors		
dbpedia-owl:academicAdvisor	508	0
dbpedia-owl:doctoralAdvisor	3,698	0
dbpedia-owl:doctoralStudent	1,791	0
dbpedia-owl:notableStudent	372	0
dbpedia-owl:influenced	2,830	0
dbpedia-owl:influencedBy	5,928	0
Keywords or topics / research area		
dbpedia-owl:knownFor	17,702	0
dbpedia-owl:notableIdea	392	0
dbpedia-owl:field	17,831	0
dbpedia-owl:significantProject	614	0

for the records in each of the data sources. We preprocess the author names to make the representation of names consistent across all data sources. The search over the index allows for slight spelling variations, the presence of only the initial of the forename, missing middle names, and a swap in the order of the fore- and surnames. The decision of whether a person record is considered to be sufficiently similar to the author of a publication record is currently based on a domain specific heuristic, and can be improved. However, the algorithm only serves as a test-bed to assess whether GND and DBpedia provide sufficiently detailed information to be used for author disambiguation.

4 Analysis and Evaluation

Analysis of GND and DBpedia. In addition to the (incomplete) list of publications of a person, GND contains additional information that could characterise a person sufficiently, including subject categories, their profession, and keywords for their publications. Unlike the GND authority file and the additional publication records, which are maintained by a library, and therefore, are structured and contain data more akin to digital libraries, DBpedia was not developed for that purpose, resulting in the required information being less readily available. In addition, the German part of DBpedia contains significantly less of the information useful for author disambiguation (see Table 1 right and Table 2).

Evaluation To determine whether the lack of more detailed information has a negative effect on the performance of author disambiguation using these reference data, we have taken the following data sets: (i) A manually created small test data set consisting of 30 of the top social scientists with a differentiated entry in GND, and DBpedia. (ii) A random subset of 250 computer scientists from

person records in DBLP with a link to the corresponding wikipedia page, and run the part of the author disambiguation algorithm that identifies the GND and DBpedia entries of an author of a publication record in Sowiport and DBLP. The precision ranging between 0.7 and 1 is encouraging (In detail: social scientists with entry in German DBpedia: 0.97; - with entry in English DBpedia: 1; - with entry in GND: 0.92; computer scientists with entry in DBpedia: 0.89 taking into account the language of the false positives; - with entry in GND: 0.7). However, the data set used here is fairly small and does not contain too many people with common names, which contribute the majority of the false positives.

5 Discussion

The analysis and evaluation of DBpedia and GND has shown that the semantic markup of the information in DBpedia is still lacking in various aspects. How much of an issue this lack of appropriately detailed information and lack of completeness really causes for tasks does not only depend on the corresponding subset of the reference data and its properties, but also on the remainder of the reference data set, and the digital library data set. This would suggest that a quality measure that assesses the suitability of the reference data set for author disambiguation should take into account the following: (i) tuple completeness, (ii) specificity of the annotation with ontologies, (iii) how much of the information is provided in form of ontologies or thesaurus or even worse literal strings, which provides an indication of the expected heterogeneity of the information across different data sets, and (iv) the number of people in the reference data set who share their names.

To bring this into context with the digital library data set, one could also determine whether and how many of the author names are shared with several person records in the reference data set. In particular in these cases, sufficiently detailed information is vital in order to be able to identify the correct person record or determine that there is no person record available for that particular person, even though there are plenty of records for people with the same name.

References

1. Ferreira, A.A., Gonçalves, M.A., Laender, A.H.F.: A brief survey of automatic methods for author name disambiguation. SIGMOD Record 41(2) (2012)
2. Herskovic, J.R.J., Tanaka, L.Y.L., Hersh, W.W., Bernstam, E.V.E.: A day in the life of PubMed: analysis of a typical day's query log. Journal of the American Medical Informatics Association : JAMIA 14(2), 212–220 (2007)
3. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S.: Dbpedia-a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal (2014)
4. Ley, M.: DBLP: some lessons learned. In: VLDB'09. pp. 1493–1500 (2009)
5. Reuther, P., Walter, B., Ley, M., Weber, A., Klink, S.: Managing the Quality of Person Names in DBLP. In: ECDL'06. pp. 508–511 (2006)
6. Smalheiser, N.R., Torvik, V.I.: Author name disambiguation. Annual review of information science and technology 43(1) (2009)