Entity Identification on the Semantic Web

Alexis Morris, Yannis Velegrakis, Paolo Bouquet

University of Trento
{morris,velgias,bouquet}@disi.unitn.eu

Abstract. In the core of every information integration and data exchange effort lies the ability to identify whether two pieces of information refer to the same real world entity. This ability is of paramount importance for all those applications and systems currently operating in the highly heterogeneous web environment. Research in data management has long ago exploited features like keys or schema constraints for dealing with that issue, but the web reality has brought new challenges. In this work we survey a number of entity disambiguation and identification techniques and tools that can be used in semantic web applications and more specifically, into an entity management system for the semantic web.

1 Introduction

Making good business decisions depends heavily not only on the amount of information available, but also on the quality of the data at hand. To successfully locate, retrieve and integrate information related to a given task, it is of paramount importance the ability to identify whether two pieces of information refer to the same real world entity. Database management and federated systems have long ago studied the specific problem by exploiting special structures such as keys or referential integrity schema constraints [1].

The advent of the web enabled the exchange of data among data sources of different organizations and individuals. Since these sources hare typically been developed by different people, at different times and with different assumptions and requirements in mind, a natural degree of heterogeneity is prevalent. This makes the problem of entity identification even harder. The same entity may be represented in different sources using different data models or meanings (semantic heterogeneity), it may be structured differently (structural heterogeneity), or it may have varying spelling values (syntactic heterogeneity). The name of a person, for instance, may be recorded in one data source using two fields for the first and last name, in a second using one field for both, while in a third it may use again one field but with the first name abbreviated. Furthermore, the fact that different organizations or individuals typically have different interests and priorities, results into a situation in which data sources model different (possibly partially overlapping) parts of the same information. For instance, a data source may store the name and date of birth of a person, while another may store only the name and the city of birth.

Entity identification can be defined as the ability of a system to accurately match semantically similar terms to the same concept. *Entity disambiguation*, a task that typically prevails entity identification, is the ability to select the most fitting categorization of a concept among a range of candidate options. In the literature the two terms are most of the time used equivalently. The problem of entity identification has been known to the data management research community for more than two decades under different names, e.g., record linkage or record matching [2], merge-purge [3], data deduplication, database hardening and very recently as reference reconciliation [4]. Most existing techniques are designed for relational systems, and may involve techniques typically met in schema matching [5] or mapping [6]. Ontologies have been used extensively to communicate the semantics of the data mostly when the schemas are limited in successfully delivering such a task. Nevertheless, neither ontologies are free from the many identification problems. Due to the size of the web, global agreement on the modeling, structure and use of the ontologies is hard to achieve. Ontologies developed by different communities are potentially diverse, and mapping techniques across ontologies have become essential [7]. Ontology integration has become one of the major challenges for the semantic web [8]. In a recent survey [9] of 25 information integration approaches that involve ontologies, one can easily note the prevalence of a need for entity identification techniques in each one of them.

An effort has recently been initiated aiming at facilitating entity identification on the Semantic Web. This effort is currently being implemented within the OKKAM [1] [10] project. The goal of OKKAM, in general, is to enable the Web of Entities, namely a virtual space where any collection of data and information about any type of web published entity, e.g. person, location, organization, event, product, etc., can be integrated into a single virtual, decentralized, open knowledge base. It provides a common global reference for every web document, application, or any other entity that has a representation on the semantic web. The success of OKKAM is of critical importance since it can offer to the Semantic Web a similar benefit to the one hypertext has offered to the Web [11]. Having entity identification as one of its core components, the success of OKKAM highly depends on the efficiency and effectiveness of the entity identification process.

In this work we survey a number of techniques and tools related to the entity identification process in the context of the semantic web. Special focus is given to approaches that use ontologies or are applicable to semantic web applications. This is the first step towards the development of the high accuracy entity identification mechanisms required in OKKAM. Critical role in all these processes plays the domain knowledge. Domain knowledge has been noted [12] as a critical tool for successfully performing tasks such as matching [5], mapping [6], evolution [13], query translation and integration [14]. The Word-Net [15] taxonomy, for instance, has been extensively used to improve the accuracy of matching methods. Of course, since dealing with semantics is a hard task, a fully automatic solution may not always be possible. Human intervention will always be

[1] http://www.okkam.org

needed, at least to verify the correctness of the generated results. Nevertheless, rich semantic models and knowledge representation reasoning techniques can significantly reduce the required human effort and allow entity identifications among data containing thousands or millions of entries [8].

The list of techniques presented here is definitely not exhaustive, mainly due to space limitations. Nevertheless, we reports all the categories we have found throughout our study. For each category we have selected one or two characteristic representative works which we report. The categories include techniques from machine learning (Section 2), lexicon/taxonomies (Section 3), similarity functions (Section 4), structural approaches (Section 5) word-sense disambiguation methods (Section 6), meta-data assisted solutions (Section 7) linguistic algorithms (Section 8) and semantic coordination 9. The goal of the current work is to provide a comprehensive picture and an understanding of how these techniques can be integrated, or maybe orchestrated, together to form a robust and efficient solution for entity identification.

2 Machine Learning

Machine learning based approaches are among the earliest works in this area. Neural Networks have been very popular in computing the semantic similarity [16]. The idea is to use a classifier to categorize attributes according to their field specifications and data values, then train a neural network to recognize similar attributes. A typical data set that has been extensively used is the DBLP[2] bibliographic collection due to its nature, size and format.

A characteristic approach of machine learning techniques for entity identification is the one of supervised learning by Han et al. [17]. The primary focus of this work is on disambiguation of names from within citation data. The authors identify name entities by the use of a "canonical" name, which is a "minimal invariant" that uniquely identifies an individual author. This technique is commonly used in libraries to overcome the problem of identity uncertainty. Despite the fact that token based methods are used, the focus is not entirely on string-based similarity calculations. Two models are used: a generative and a discriminative. The generative model results in new sample data through the use of a Naive Bayes algorithm. A support vector machine adapted for multi-class classification is used as the discriminative model. To illustrate in more details the specific mechanism we provide a high level description of the performed steps on the DBLP data set. The Naive Bayes Model is initially setup by computing a probability based on a group of citation entries that have been parsed by regular expressions to determine a sample set. Each entry is analyzed to see if the maximum posterior probability indicates that the author entry is the author of a paper. The processing of the author information is used in future iterations as prior knowledge or as training data. The Bayes rule is used to compute which author wrote which paper. In the sequel, the training data is used to compute the probability of writing a paper in the future with some other authors. To do so, sub-probabilities are calculated, such as: the probability of writing a paper

[2] http://www.informatik.uni-trier.de/ ley/db

alone, with co-authors in general, with previously seen co-authors, with unseen co-authors, etc. These individual probabilities are then combined with the terms found in the citations to keep track of author interests versus co-author patterns. In a different process, a support vector machine approach is used in order to classify a citation to the nearest author. Citations are given a vector containing features of the author information and other parameters. Finally, a decision function is computed and the classification features are ranked for future interactions.

Although use of a machine learning approach is generic, the above ideas can be easily extended to take into consideration similarities between terms and concepts as provided by ontologies or taxonomies. For instance, WordNet has been suggested for that purpose [17]. Clustering based on that similarity can then be applied. The advantage of the machine learning approaches is that they have a high precision and recall, but on the other hand, they require training which may not always be easy. Furthermore, as can be noticed above, the analysis performed on the data, it is not only based on similarities, but takes into consideration behavioral characteristics, i.e., with what persons has an author previously collaborated. The findings can then be used to predict future behaviour.

3 Lexicon/Taxonomy

A commonly used approach to quest of discovering the semantics of entities is to use a taxonomy structure, an ontology, a concept map, or a lexicon, i.e., a data dictionary. The Adapted Lesk algorithm [18] is one of those proposed for word sense disambiguation. It works by computing how semantically close two words are. To disambiguate and assign the right meaning to a word, context information is used. The notion of *gloss overlap* is also used as a measure, where "gloss" represents the definition of a word, i.e., glossary, and the corresponding "overlap" is a measure of the number of words that are common to different gloss-groups. The approach is based on taxonomy concept hierarchies such as WordNet. The gloss calculation is based on a weighted scoring method. A central issue in lexicon-based approaches is the definition of the similarity between lexicon terms. Among the different metrics that can be used for such a purpose is the length of the path between the two words in the hierarchy, the kind of edges that exist in such a path, the context information, or a combination of the above. Another similarity measure is the ratio of the amount of information needed to state the commonality of the two concepts over the amount of information needed to describe them. To decide whether two concepts correspond to the same real world entity or not, one needs to compute the similarity between their respective concepts and compare it to the similarity of other concepts. The decision is based on some predefined cut-off value.

The advantages of using taxonomies is that they may be leveraged for their capacity of domain information for different purposes. Furthermore, taxonomies facilitate semi-automatic solutions. On the other hand, the limitation of taxonomy based approaches is that they require a shared taxonomy to be always available and a domain expert to tune the cut-off values on which the decisions on entity identification are based.

4 Similarity Functions

The most commonly used approach in entity identification is probably the use of similarity functions. A similarity function is a function that computes a score based on how many components two entities have in common or not. Popular functions are the vector model [19], distance measurements such as those proposed in information theory [20] or those based on relational schemes [21]. A system for global schema generation and integration based on ontologies has recently been developed [22]. The system integrates local ontologies into a global counterpart. It performs similarity inference using description logic reasoning. This reasoning provides a mean of disambiguation through equality, specialization, overlapping, and disjoint relationships.

Seeker [23] is one of the largest, in terms of scale, semantic tagging efforts to date. It is a text analysis engine used mainly for web annotation. It performs automatic entity disambiguation using a technique called Taxonomy Based Disambiguation. It employees similarity functions, along with machine learning techniques, to calculate the semantic value of a given word and generate semantically meaningful tags.

For a similar goal, i.e., semantic annotation, similarity functions have been used [24] to disambiguate author names from citation data using an "ontology category utility". In the particular approach, authors are initially represented as clusters of their published work. Key terms are extracted from the titles and the abstracts based on the concepts that exist in a domain ontology. Authors are then compared to each other, based on four types of similarities: exact-match, taxonomy similarity, subsumption similarity, and relation similarity. The results of each comparison is a set of four numbers, known as the the 4 category utilities. The four category utilities are then summed up and the total represents the final score that determined whether two authors are actually the same person or not.

The advantage of similarity functions is that they are easy to formalize and they can be easily tuned to adapt to the nature of the data at hand in order to maximize recall and precision.

5 Structure Exploitation

A great part of the semantics of every piece of information is stored in its structure. This is why good data modeling is considered as one of the most critical tasks in data management. This extends to ontological data as well. One can analyze the structure of the ontology concepts in order to determine whether two concepts refer to the same real world entity or a similar one. Methods that have been developed for semi-structured data [25] can easily be adapted to apply in the case of ontologies. A recent study [26] suggests three different measures for comparing concepts, namely graph matching, filter-to-one descriptions and probabilistic measures. The same work also describes a structural entity matching approach based on tree-edit distance, which is intended to be used as a first step towards the building of more general integration approaches. It aims at creating a computational model that assesses semantic similarity among entity classes from different and independent ontologies without constructing a-priory

a shared ontology. To compute the structural distance, three specific matchings are taken into consideration, i.e., word matching, feature matching, and neighborhood matching. The results of each one are combined into a weighted-sum similarity function which also incorporates the depth of the taxonomy or the ontology tree. For word matching the authors check the words in synonym sets by comparing the number of similar and different words. This approach is able to find the highest degree of similar terms between different synonym sets and gives an estimate of similarity. Feature matching refers to the process of determining the distance between distinguishing features. These features are lexicographic, and involve string matching over words in synonym sets. Finally, for semantic-neighbourhood matching, entity classes are grouped into a semantic cluster according to the synonym set or feature matching. This type of matching is based on a quantity defined as the cardinality of the neighbourhoods over the cardinality of their intersection.

The accuracy of structural exploitation can be highly improved through knowledge of distinguishing features of certain types of entities. For instance, knowing that the date and place of birth, along with the name are distinguishing features of a person, one can conclude that to entities refer to the same person if they agree on these distinguishing attributes even if they differ in the rest of their structure.

6 Word-sense Disambiguation

One the most important tasks in entity identification is the word sense disambiguation. Naturally, one cannot expect that the same word has always the same meaning in every web applications. This observation has led to a number of studies on finding ways to make sense of words based on the context in which they appear. Id-Rank [27] is one such approach. The target domain is the multi-domain news archives of multiple news agencies. It is designed similarly to the Page-Rank algorithm, but uses a news metadata ontology and a natural language processing engine, along with a heuristic/deductive database techniques. It is based on the notion of semantic coherence, i.e., common appearances of entities in certain contexts, and trends, i.e., important references to a particular entity or event.

In analogy to Page-Rank that values a page based on its incoming reference links, Id-Rank values the meaning of a word based on the number of news documents in which the word appears, and is mainly used to rank candidate entities. A semantic network is formed from the entities in the candidate set. An edge between two entities means that the entities have at some point appeared in the same context. Edges have weights. An equation incorporates these weights in a computation process in order to come up with a final score which will be used to generate the final ranking of the candidate entities similar to an entity at hand.

The advantage of word-sense disambiguation approaches is that they can find similarities between entities modeled at different times and different fields, since they do not require a common vocabulary. Nevertheless, due to the semantic information that is involved in the task, a verification process by a domain expert may often be necessary.

7 Metadata Information

The importance of metadata has already been recognized in many different fields [28]. Specifically for entity identification, metadata can play a significant role. The term metadata refers to any kind of information that is not considered part of the data but is related to it and helps in better communicating its semantics. Metadata may include provenance information, language or quality information, i.e., time, accuracy, authority, etc.

Ontologies are one of the main tools used to communicate metadata information. Mappings between different ontologies can be used to achieve interoperability between different ontologies [29]. This can be either on one-to-one basis, i.e., a P2P style, or in a style similar to information integration systems. In the latter case, a global ontology is created and then mappings are introduced to associate the terms and concepts of the global ontology to those of the individual local ontologies. Mappings are typically logical expression. This provides the additional advantage that inference and integration systems can use, for instance, description logic rules to find subsumption relationships, inconsistencies, and provide query optimization services.

Additional kinds of metadata, such as language information, can be used to achieve the linking of different values that may represent the same real world entity but look different just because they are expressed in different languages.

8 Linguistic Analysis

Social reasons are an important form of heterogeneity. Different people may use different words or expressions to refer to the same entity. In such cases, linguistic analysis technologies can be proved highly beneficial in disambiguating words or expressions in the data or in the user queries. A successful application of linguistic techniques is the Linguistic Combination System (LCS) [30]. First, LCS maps terms of two ontologies. The results are then provided as input to matchers which produce a similarity cube that is then aggregated into a similarity matrix and used for richer mapping discovery. The approach makes use of similarity functions, but proposes its own linguistic aggregation operation to determine matches. It involves several phases, proceeding via matching functions for the two candidate ontologies, as well as aggregation functions for deducing the similarity possibilities, and heuristics to discover mappings based on the prior phases of execution. The approach begins with ontology matchers for Name, Name path, Taxonomy, Domain and Range, and Mother-concept. The initial matcher, the Name matcher, provides the first similarity matrix that is used by the other matchers as input. The Linguistic Aggregation Operators (LAO) are then chosen by a user and are applied for every entity pair (x,y). These are operators for Max, Min, Avg, Most, Alh (at least half), and Amap (as many as possible). They are used to group the match results as needed based on which matchers are satisfied or not. Once the matchers are combined to form the similarity cube, the LAOs are used to further combine these into a similarity matrix for selecting match candidates.

9 Semantic Coordination

Semantic coordination, namely the problem of finding an agreement on the meaning of heterogeneous semantic models, is one of the key issues in the development of the Semantic Web. A new algorithm has recently been proposed [31] for the discovering of semantic mappings across hierarchical classifications based on semantic coordination. It employees lexical structure and semantic similarity. The specific approach shifts the problem of semantic coordination from the problem of computing linguistic or structural similarities (what most other proposed approaches do) to the problem of deducing relations between sets of logical formulae that represent the meaning of concepts belonging to different models. The approach is based on the intuition that there is an essential conceptual difference between coordinating generic abstract structures (e.g., arbitrary labelled graphs) and coordinating structures whose labels are taken from the language spoken by the community of their users. An important conclusion of the above work is that the intended use of the data is a very important factor for entity identification, equivalent or even more important that the actual structure of the data.

10 Discussion and Conclusions

We have carried out a study on the different approaches, techniques and tools that can be used for entity identification on the semantic web. Entity identification is not a new problem and many studies have already been performed and a number of interesting surveys have already been published [32]. However, these works have concentrated their effort on database management techniques. Their main interest is on duplicate record detection on databases with millions of tuples. Here we put effort more on the semantic aspect of the problem. We chose to present approaches that are either based on semantic web applications and ontologies, or can be used for entity identification on ontological data. The motivation for our work was an advanced entity matching mechanism that is currently under development within the OKKAM project, a project aiming to enable the web of entities by providing an infrastructure that can be used to assign identifiers to every entity available on the semantic web.

The conclusion of our study is that there is no silver bullet. No single technique can offer a service that performs well in all the situations. Despite the depth and breath of the existing techniques we believe that there is enough space for optimization and improvement mainly in two directions. The first is the development of methods that exploit the results of different approaches and combine them to reach one single decision. The second direction is the incorporation to the existing techniques of domain knowledge.

Another observation that we have made is that all the existing techniques operate on the assumption of static, or relatively static, data. The evolving nature of the data and schemas has not been taken into consideration. The modern web is a very volatile environment since millions of users are not only accessing the data but are also modifying it. Thus, mechanisms to easily and quickly adapt to new requirements and new data are needed [33]. Furthermore, recent advances

on sensor networks and other technologies, have introduced numerous streaming data sources. In those cases, many of the presented entity identification techniques will be hard to apply [34]. The reason is that the streaming nature does not allow for expensive computations and for large amount of data to be kept in memory or on the hard disk, thus, the existing methods will have to be adapted and new one may have to be invented.

Acknowledgements: This work has been partially funded by the EU grant ICT-215032.

References

1. Arens, Y., Chee, C.Y., Hsu, C., Knoblock, C.A.: Retrieving and Integrating Data from Multiple Information Sources. JICIS **2** (1993) 127–158
2. Newcombe, H.B.: Record linking: The design of efficient systems for linking records into individual and family histories. American Journal of Human Genetics **19** (1967) 335–359
3. Hernandez, M.A., Stolfo, S.J.: Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. Data Mining and Knowledge Discovery **2** (1998) 9–37
4. Dong, X., Halevy, A.Y., Madhavan, J.: Reference Reconciliation in Complex Information Spaces. In: SIGMOD Conference. (2005) 85–96
5. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDB Journal **10** (2001) 334–350
6. Popa, L., Velegrakis, Y., Miller, R.J., Hernandez, M.A., Fagin, R.: Translating Web Data. In: VLDB. (2002) 598–609
7. Choi, N., Song, I., Han, H.: A survey on ontology mapping. SIGMOD Rec. **35** (2006) 34–41
8. Noy, N.F.: Semantic integration: a survey of ontology-based approaches. SIGMOD Record **33** (2004) 65–70
9. Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S.: Ontology-based integration of information — a survey of existing approaches. In: IJCAI–01 Workshop: Ontologies and Information Sharing. (2001) 108–117
10. Palpanas, T., Chaudhry, J., Andritsos, P., Velegrakis, Y.: Entity Data Management in OKKAM. In: SWAP DEXA Workshop. (2008) 729–733
11. Bouquet, P., Stoermer, H., Bazzanella, B.: An Entity Name System (ENS) for the Semantic Web. In: ESWC. (2008) 258–272
12. Doan, A., Halevy, A.Y.: Semantic-integration research in the database community. AI Mag. **26** (2005) 83–94
13. Velegrakis, Y., Miller, R.J., Popa, L.: Mapping Adaptation under Evolving Schemas. In: VLDB. (2003) 584–595
14. Pinto, H., Prez, A., Martins, J.: Some Issues on Ontology Integration. In: IJCAI-99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends. (1999)
15. Niles, I., Pease, A.: Linking lexicons and ontologies: Mapping WordNet to the suggested upper merged ontology. In: Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE '03), Las Vegas, Nevada. (2003)

16. Li, W., Clifton, C.: Semantic Integration in Heterogeneous Databases Using Neural Networks. In: VLDB. (1994) 1–12
17. Han, H., Giles, L., Zha, H., Li, C., Tsioutsiouliklis, K.: Two supervised learning approaches for name disambiguation in author citations. In: JCDL. (2004) 296–305
18. Patwardhan, S., Banerjee, S., Pedersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: In International Conference on Intelligent Text Processing and Computational Linguistics. (2003) 241–257
19. Frakes, W.B., Baeza-Yates, R.A.: Information Retrieval: Data Structures & Algorithms. Prentice-Hall (1992)
20. Lin, D.: An Information-Theoretic Definition of Similarity. In: ICML. (1998) 296–304
21. Brauner, D.F., Intrator, C., Freitas, J.C., Casanova, M.A.: An Instance-based Approach for Matching Export Schemas of Geographical Database Web Services. In: GeoInfo. (2007) 109–120
22. Hakimpour, F., Geppertb, A.: Resolving semantic heterogeneity in schema integration. In: FOIS. (2001) 297–308
23. Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R.: SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation. In: WWW. (2003) 178–186
24. Park, Y., Kim, J.: OnCU system: ontology-based category utility approach for author name disambiguation. In: ICUIMC. (2008) 63–68
25. Madhavan, J., Bernstein, P.A., Rahm, E.: Generic schema matching with cupid. In: VLDB. (2001) 49–58
26. Rodriguez, M.A., Egenhofer, M.J.: Determining Semantic Similarity among Entity Classes from Different Ontologies. TKDE **15** (2003) 442–456
27. Garca, N.F., del Toro, J.M.B., Snchez, L., Bernardi, A.: IdentityRank: Named Entity Disambiguation in the Context of the NEWS Project. In: ESWC. (2007) 640–654
28. Srivastava, D., Velegrakis, Y.: Intensional Associations between Data and Metadata. In: SIGMOD. (2007) 401–412
29. Kwon, J., Jeong, D., Lee, L., Baik, D.: Intelligent semantic concept mapping for semantic query rewriting/optimization in ontology-based information integration system. International Journal of Software Engineering and Knowledge Engineering **14** (2004) 519–542
30. Ji, Q., Liu, W., Qi, G., Bell, D.A.: LCS: A Linguistic Combination System for Ontology Matching. In: KSEM. (2006) 176–189
31. Bouquet, P., Serafini, L., Zanobini, S.: Semantic Coordination: A New Approach and an Application. (2003) 130–145
32. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate Record Detection: A Survey. TKDE **19** (2007) 1–16
33. Velegrakis, Y.: On the Importance of Updates in Information Integration and Data Exchange Systems. In: DBISP2P. (2008)
34. Tantono, F.I., Manerikar, N., Palpanas, T.: Efficiently discovering recent frequent items in data streams. In: SSDBM. (2008) 222–239