Entity Disambiguation for Wild Big Data Using Multi-Level Clustering

Jennifer Sleeman

Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore. MD 21250 USA
jsleem1@cs.umbc.edu

Abstract. When RDF instances represent the same entity they are said to corefer. For example, two nodes from different RDF graphs[1] both refer to same individual, musical artist James Brown. Disambiguating entities is essential for knowledge base population and other tasks that result in integration or linking of data. Often however, entity instance data originates from different sources and can be represented using different schemas or ontologies. In the age of Big Data, data can have other characteristics such originating from sources which are schema-less or without ontological structure. Our work involves researching new ways to process this type of data in order to perform entity disambiguation. Our approach uses multi-level clustering and includes fine-grained entity type recognition, contextualization of entities, online processing of which can be supported by a parallel architecture.

Introduction

Often when performing knowledge base population, entities that exist in the knowledge base need to be matched to entities from newly acquired data. After matching entities, the knowledge base can be further enriched with new information. This matching of entities is typically called entity disambiguation (ED) or coreference resolution when performed without a knowledge base [15]. Early work related to record linkage [6] was foundational to this concept of entity similarity. Though there is a significant amount of research in this area including methods which are supervised and unsupervised, these approaches tend to make assumptions that do not hold for big data.

Existing research tends to assume a static batch of data, ignoring the streaming, temporal aspects. It assumes that the schemas or ontologies are available and complete. Often issues such as heterogeneity and volume are not considered. However, big data applications tend to include unalignable data from multiple sources and often have schemas or ontologies that are absent or insufficient. We define these characteristics in terms of 'Wild Big Data' (WBD) [21] and describe how these characteristics challenge the disambiguation process. Our work specifically addresses these characteristics with an approach that could be used to perform ED for WBD.

[1] http://dbpedia.org/resource/James_Brown and http://musicbrainz.org/artist/20ff3303-4fe2-4a47-a1b6-291e26aa3438#

Objective

The objective of this research is to perform ED given the data is large in volume, potentially schema-less, multi-sourced and temporal by nature. We want to answer questions such as, how do we perform ED in a big data setting, can we efficiently distribute the task of ED without a loss in precision, how do we account for data that is changing over time, how do we process semantic graphs given they may not have an associated schema or ontology, and finally how do we process this data given it originates from different sources with potentially unalignable vocabularies. These questions are important to answer because they are real problems in big data applications [25].

Motivation

Big data is a growing area of research and offers many challenges for ED [2, 11]. The main motivation of this work is the need for ED that supports data with big data characteristics. This includes data originating from different sources which contain different types of entities at different levels of granularity, data that may not have a schema or ontology, and data that changes over time. This sort of data at big data volumes complicates the ED process.

Companies, organizations and government entities are sharing more data and acquiring more data from other sources to gain new insight and knowledge [8]. Often the combination of sources, such as social media, news and other types of sources can provide more insight into topics than a single source.

As is evident by efforts related to Linked Open Data (LOD) [17], interoperability among different data sources is of growing importance and essential for sharing data. As more data is made available for sharing, the need for aligning schemas/ontologies is increasing.

Knowledge bases typically contain entities, facts about the entities and links between entities. As new data is made available over time, these knowledge bases require ways to manage new information such as adding entities, links and new attributes pertaining to the entities. There is a need to also alter existing information such that information that becomes invalid over time is adjusted. For example, a link may become invalid or an attribute may prove to be incorrectly assigned to an entity.

Challenges and Opportunities

By exploring how to perform ED for big data, we will offer a strong contribution to this area as previous research has only focused on various parts of this problem.

Regarding the LOD [17], interoperability is a real challenge, particularly because vocabularies are not always alignable. For example, *address* in one vocabulary could mean *street address* alone and in another it could include *city*, *state* and *zip code*. We explored this problem in more depth in our previous work [18]. LOD attempts to provide a way for data providers to link their data into the cloud. However data may not always be made available as LOD, and in order for an application to perform ED, this alignment becomes essential.

With unstructured text, one can use natural language processing to acquire various facts related to entities found in the text. With RDF data, an ontology can often be used to develop an understanding of the data. However, when data is semi-structured such as RDF or JSON and no such ontology or schema is present, disambiguating entities becomes problematic. Making sense of these large data extractions becomes a real issue.

Knowledge bases naturally change over time, however it is a challenge to enrich the knowledge base over time while at the same time reducing errors in previously asserted facts. Algorithms used to perform ED are typically developed for static data. Incremental updates and changes are harder to incorporate. However, this is precisely what is needed as often big data applications are producing data on a periodic basis. If one is developing a knowledge base where facts are changing over time, the ED algorithm must accommodate these changes in a way that does not require the algorithm to reprocess all potential matches given new information.

Volume requires that the algorithm can be distributed in such a way that work could be performed in parallel. Again ED algorithms do not typically assume data in terms of the volume that is present with big data applications. However, since ED algorithms have typically $O(n^2)$ complexity, distributing the algorithm would be necessary for such large volumes of data.

Recent research which has addressed big data ED has primarily been in the natural language processing domain. For example, a number of researchers [4, 13, 16] have explored using MapReduce for pairwise document similarity. However, they are primarily focused on the volume characteristic. Work by Araujo et al. [1] tackled the problem of working with heterogeneous data but they worked with sources where the vocabularies were alignable. Work by Hogan et al. [9] addresses this problem of performing ED for large, heterogeneous data. However, they assume they have access to the ontologies used and they assume they can make use of owl:sameAs semantics (which isn't always present). The hard problem of trying to understand data absent knowledge of how it is structured has not been thoroughly addressed in previous research.

Proposed Approach

We are developing a multi-level clustering approach that includes one level of topic modeling and a second level of clustering using our own custom algorithm. This approach makes big data ED more tractable. Our research makes three major research contributions that work together to achieve an effective approach for performing online ED.

Research Contribution: Fine-grained Entity Type Recognition: If we consider identifying traits of an entity, at the highest level of identification, entities are defined by types, for example "Person", "Football Player", "Baseball Stadium", etc. With TAC [2] there are just three types used (PER, ORG, GEP) used, with DBpedia there are fewer than 1000 types, and tens of thousands of types in Yago. Given a WBD data set, data can contain a mix of entities,

[2] http://www.nist.gov/tac

can be composed of many different types, such as a person, a sports player, a team member, and can be defined by types that are defined at different levels of granularity. For example, “Person” is at a much higher level than “Football Player”. Often type information is not available, to get around this problem, we have proposed a solution [21] based on topic modeling that enables us to define types when type information is not present.

Research Contribution: Multi-dimensional Clustering: We are developing a clustering algorithm that performs ED based on multiple dimensions. This algorithm would be applied to the coarse clusters generated from the fine-grained entity type recognition. The complexity of clustering algorithms can range from $O\left(n^2\right)$ to $O\left(n^3\right)$, so a key aspect of this work is that it supports parallelism.

Research Contribution: Incremental Online modeling to support Temporal Change: Our work includes knowledge base (KB) population. When entities are assessed as similar, the information in the KB is merged with the information contained in the newly recognized matched entity instance. However, similarity is usually associated with some level of probability. As more data is acquired over time, previous assertions may prove to have a lower probability than previously asserted.

Relationship with State of the art

As it relates to coreference resolution the following work [12, 1, 24] would be considered state of the art and is comparable to our work. In the NLP domain, a number of researchers have focused on scalable entity coreference using MapReduce [4, 13, 16].

As it relates to type identification, work by Ma et al. [10] presents a similar problem, whereby type information is missing. This work builds clusters that represent entity types based on both schema and non-schema features. Paulheim et al. [14] also address the problem of identifying type information when it is non-existent and they also use their approach to validate existing type definitions. They take advantage of existing links between instances and assume that instances of the same types should have similar relations. They acquire this understanding by examining the statistical distribution for each link.

As it relates to candidate selection, the following work [23, 15] would be considered state of the art and comparable to our work.

Implementation of Proposed Approach

We will implement our approach as a software system by which it could be used to perform ED for wild big data. We will demonstrate the effectiveness of this approach by launching it in a parallel environment processing wild big data. We will use benchmarks to convey the overall performance of the system as it compares to other systems that are not necessarily addressing the wild big data aspects. We anticipate ED scores that have slightly lower precision but we expect to see better computing performance as we scale the number of entities in our system to big data sizes, since our approach is developed to be amenable to a Hadoop-like architecture.

Current Implementation

For the first level of clustering we use Latent Dirichlet Allocation (LDA) [3] topic modeling, to form coarse clusters of entities based on their fine-grained entity types. We use LDA to map unknown entities to known entity types to predict the unknown entity types. We shared preliminary results of this effort in our previous work [21]. Table 1 shows our latest results that include experiments using DBpedia data where we show accuracy given we found all types and accuracy given we missed on 1 type but found the others. Figure 1 also shows another experiment where we measured precision at N where given N predictions we found all of the types for a particular entity. This approach offers two benefits, it results in overlapping clusters based on entity types improving recall and it does not require knowledge of the schema or ontology of the data. The only requirement is that there is a knowledge base of entity types that can be used as a source for associating entity types to unknown entities.

We have performed research related to ED of people in our early work [19] where we experimented with combining rules and supervised classification, however when ED is performed on entities of different types in combination with different data sources, the ED process is more difficult. Often recognizing the types of entities and then performing ED among specific types can reduce this problem, however, when the data sets are large to the scale of big data problems, even recognizing these types reduces the problem to intractable sized subproblems. For this reason, we are building a custom clustering algorithm for the second level of clustering. This work is still in-process.

We have performed preliminary work [20] with hierarchical clustering and did not find this to be a viable solution. Our current work clusters based on a number of features such as distance measures, co-occurrences, graph-based properties, and statistical distributions. Distinctive to our work, we also incorporate context which we derive from our topic model. Entity context provides additional information about an entity that is not necessarily acquired from the associated predicates for that entity. We are also currently performing preliminary experiments related to contextualizing entities.

Current Limitations

Since our approach is a two-level approach, errors from the first level of clustering could propagate to the second level. We look to overcome this problem by generating a model that both levels of clustering would use, however a resolution to this problem is still under investigation.

This approach is currently limited to graph-based data. There is a lot of unstructured text and it would be advantageous for our system to be able to convert unstructured text to graph-based structures. In addition, in order for our approach to work with data that is truly “wild“, we require access to a knowledge base that is rich with fine-grained entity types. The richness of the knowledge base and its representation of the data to be processed directly influence how well our approach will perform. For example, if our knowledge base has very little information related to car accidents and we are processing entities from a data source related to car accidents, we will under-perform when recognizing the fine-grained entity types which consequently will negatively impact our ED algorithm.

Empirical Evaluation Methodology

Since there are multiple parts to our approach, we intend to evaluate the various parts in addition to how well the parts work together to perform ED.

Hypotheses

1. By using a multi-level clustering approach we can perform ED for wild big data and achieve F-measure rates that are close to those of other ED algorithms that are not processing wild big data.
2. Fine-grained entity type recognition as a first level of clustering is a competitive approach to performing candidate selection.
3. Our approach will be scalable such that it is comparable with other methods that perform ED in parallel.
4. By performing ED online, we can reduce the number of errors in our KB.

General Strategy

Our general approach for evaluation is to evaluate our first level of clustering, the fine-grained entity type recognition work in isolation of ED. We will then perform experiments related to contextualizing entities, performing ED both from a scalability and accuracy perspective, and finally online KB improvements.

Benchmarks We will use data sets that we are able to easily establish ground truth for, such as DBpedia and Freebase. However, we will also use Big Data datasets and we may use unstructured data sets that are processed by an OpenIE [5] system resulting in triple-based information.

Our goal with the fine-grained entity type recognition work is to be able to identify all entity types that are assigned to gold standard entities. We also will try to identify incorrectly used and missing entity types. We will perform experiments which benchmark our first level of clustering with other candidate selection methods and will be benchmarked against an existing type identification approach [14].

With our second level of clustering we hope to demonstrate that contextualization of entities improves performance. We also plan to compare our ED with others from an accuracy standpoint and from a complexity standpoint. We will benchmark how well we scale in a parallel environment compared to other parallel ED approaches.

One feasible approach for evaluating the ED method is to use data from the LOD and remove links then compare our results with the unlinked data to the data that is linked [7]. We will also explore the Instance Matching Benchmark

[3] for evaluation and benchmarking. Another benchmark that is more recent is SPIMBench [4] which provides test cases for entity matching and evaluation metrics, and supports testing scalability. Finally we will show how a KB with online temporal changes can reduce errors over time. We will prove this by taking an offline KB and comparing it to our online version.

Metrics For our evaluation we will use the standard F-measure metric. For evaluating our clusters, we will likely use standard clustering metrics such as measuring purity.

$Precision = \frac{TruePositive}{TruePositive+FalsePositive}$
$Recall = \frac{TruePositive}{TruePositive+FalseNegative}$
$F-measure = 2 * \frac{Precision*Recall}{Precision+Recall}$

Current State of Evaluation

Our early work [22] shows our evaluation of identifying fine-grained entity types using an entropy-based approach. We now use a topic modeling approach and have performed preliminary evaluation of this work [21]. We also include in Table 1 our latest evaluation. This evaluation is based on DBpedia 6000 randomly selected entities and 176 types used to build the model. We used 350 separately randomly selected entities that are of type *Creative Works*, type *Place*, and type *Organization*, as these had the highest representation among the training set. We measured how often we were able to recognize all types associated with each entity as defined by DBpedia. We are also in the process of a comprehensive evaluation for this work. We are currently developing our custom clustering algorithm and will plan to evaluate this work soon. We performed preliminary experiments with an online KB where we reduced the errors by 70% by updating the KB over time.

Table 1: Fine-Grained Entity Type Accuracy

Test	Avg Num Types	Accuracy (0 Types Missed)	Accuracy (1 Type Missed)
CreativeWork	6	.76	.91
Place	7	.60	.67
Organization	9	.74	.77

Lessons Learned, Open Issues, and Future Directions

One of our challenges is finding the data we need to properly evaluate our approach. Since we are proposing a system that works with Big Data scale datasets, our evaluations will be harder to achieve.

A second challenge is comparing and benchmarking our work against others. Since our approach addresses problems that may overlap with other research

[3] http://islab.dico.unimi.it/iimb/
[4] http://www.ics.forth.gr/isl/spimbench/index.html

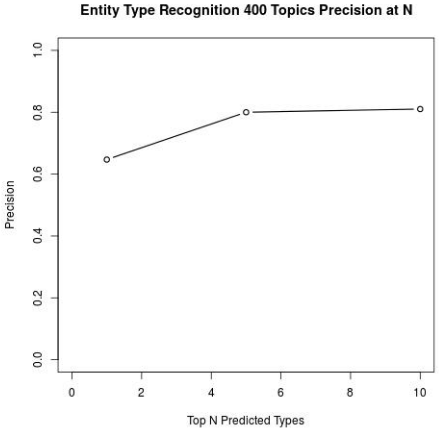

Fig. 1: Fine-Grained Entity Type Precision at N

but isn't exactly the same, we will need to benchmark parts of our system with other research.

From our previous experiments when evaluating mappings of entity types from one data source to another we learned that since there will not always be a direct mapping, we will need to have supporting heuristics which makes the evaluation process harder to achieve. For example mapping between Freebase and DBpedia is not always possible, often because types defined in one knowledge base just do not exist in the other.

Acknowledgments

The author would like to thank her advisor, Dr. Tim Finin, and Dr. Anupam Joshi.

References

1. Araujo, S., Tran, D., DeVries, A., Hidders, J., Schwabe, D.: Serimi: Class-based disambiguation for effective instance matching over heterogeneous web data. In: WebDB. pp. 25–30 (2012)
2. Beheshti, S.M.R., Venugopal, S., Ryu, S.H., Benatallah, B., Wang, W.: Big data and cross-document coreference resolution: Current state and future opportunities. arXiv preprint arXiv:1311.3987 (2013)
3. Blei, D.M.: Probabilistic topic models. Communications of the ACM 55(4), 77–84 (2012)
4. Elsayed, T., Lin, J., Oard, D.W.: Pairwise document similarity in large collections with mapreduce. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. pp. 265–268. Association for Computational Linguistics (2008)
5. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Communications of the ACM 51(12), 68–74 (2008)

6. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. Journal of the American Statistical Association 64(328), 1183–1210 (1969)
7. Ferrara, A., Montanelli, S., Noessner, J., Stuckenschmidt, H.: Benchmarking matching applications on the semantic web. In: The Semanic Web: Research and Applications, pp. 108–122. Springer (2011)
8. Franks, B.: Taming the big data tidal wave: Finding Opportunities in Huge data streams with advanced Analytics, vol. 56. John Wiley & Sons (2012)
9. Hogan, A., Zimmermann, A., Umbrich, J., Polleres, A., Decker, S.: Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. Web Semantics: Science, Services and Agents on the World Wide Web 10, 76–110 (2012)
10. Ma, Y., Tran, T., Bicer, V.: Typifier: Inferring the type semantics of structured data. In: Data Engineering (ICDE), 2013 IEEE 29th Inter. Conf. on. pp. 206–217. IEEE (2013)
11. McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D., Barton, D.: Big data. The management revolution. Harvard Bus Rev 90(10), 61–67 (2012)
12. Nikolov, A., Uren, V., Motta, E., Roeck, A.: Overcoming schema heterogeneity between linked semantic repositories to improve coreference resolution. In: Proc. 4th Asian Conf. on the Semantic Web. vol. 5926, pp. 332–346 (December 2009)
13. Pantel, P., Crestan, E., Borkovsky, A., Popescu, A.M., Vyas, V.: Web-scale distributional similarity and entity set expansion. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. pp. 938–947. Association for Computational Linguistics (2009)
14. Paulheim, H., Bizer, C.: Type inference on noisy rdf data. In: International Semantic Web Conference (2013)
15. Rao, D., McNamee, P., Dredze, M.: Entity linking: Finding extracted entities in a knowledge base. In: Multi-source, Multilingual Information Extraction and Summarization, pp. 93–115. Springer (2013)
16. Sarmento, L., Kehlenbeck, A., Oliveira, E., Ungar, L.: An approach to web-scale named-entity disambiguation. In: Machine Learning and Data Mining in Pattern Recognition, pp. 689–703. Springer (2009)
17. Schmachtenberg, M., Bizer, C., Paulheim, H.: State of the LOD cloud. http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/ (2014)
18. Sleeman, J., Alonso, R., Li, H., Pope, A., Badia, A.: Opaque attribute alignment. In: Proc. 3rd Int. Workshop on Data Engineering Meets the Semantic Web (2012)
19. Sleeman, J., Finin, T.: Computing foaf co-reference relations with rules and machine learning. In: The Third Int. Workshop on Social Data on the Web. ISWC (November 2010)
20. Sleeman, J., Finin, T.: Cluster-based instance consolidation for subsequent matching. Knowledge Extraction and Consolidation from Social Media p. 13 (2012)
21. Sleeman, J., Finin, T.: Taming wild big data. In: Symposium on Natural Language Access to Big Data (2014)
22. Sleeman, J., Finin, T., Joshi, A.: Entity type recognition for heterogeneous semantic graphs. In: AI Magazine. vol. 36, pp. 75–86. AAAI Press (March 2105)
23. Song, D., Heflin, J.: Automatically generating data linkages using a domain-independent candidate selection approach. In: Int. Semantic Web Conf. (2011)
24. Song, D., Heflin, J.: Domain-independent entity coreference for linking ontology instances. Journal of Data and Information Quality (JDIQ) 4(2), 7 (2013)
25. Suchanek, F., Weikum, G.: Knowledge harvesting in the big-data era. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. pp. 933–938. ACM (2013)