Jennifer Sleeman and Tim Finin, Taming Wild Big Data, AAAI Fall
Symposium on Natural Language Access to Big Data, Nov. 2014.

Taming Wild Big Data

Jennifer Sleeman and **Tim Finin**
Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore. MD 21250 USA
{jsleem1,finin}@cs.umbc.edu

Abstract

Wild Big Data (WBD) is data that is hard to extract, understand, and use due to its heterogeneous nature and volume. It typically comes without a schema, is obtained from multiple sources and provides a challenge for information extraction and integration. We describe a way to subduing WBD that uses techniques and resources that are popular for processing natural language text. The approach is applicable to data that is presented as a graph of objects and relations between them and to tabular data that can be transformed into such a graph. We start by applying topic models to contextualize the data and then use the results to identify the potential types of the graph's nodes by mapping them to known types found in large open ontologies such as Freebase, and DBpedia. The results allow us to assemble coarse clusters of objects that can then be used to interpret the link and perform entity disambiguation and record linking.

Introduction

Big Data in recent years has received a lot of attention with expectations that it will only keep growing (Chen, Chiang, and Storey 2012; McAfee et al. 2012; Franks 2012; Wu et al. 2014). Where Mcafee et al. (McAfee et al. 2012) reported in 2012 2.5 exabytes of data created daily. However, there are problems that are still unresolved as it relates to the V's of Big Data (Hendler 2013; McAfee et al. 2012; Dong and Srivastava 2013; Hitzler and Janowicz 2013).

Though volume can be beneficial for machine learning, it presents a problem for processing the actual data and increases the likelihood of error (Mayer-Schönberger and Cukier 2013). Velocity or the rate by which data is received also presents a problem because systems need to account for a continuous stream of data. Variety or diversity of data sources implies data can originate from sources of different domains or within the same domain data but originate from different subdomains (Franks 2012). This results in data using different schemas or representation (if any at all), and data expressed in different formats such as unstructured text, images and video. Combining data of different types and from different sources is not an easy task and presents yet another opportunity for error (Mayer-Schönberger and Cukier 2013). Data is often generated by machines and companies are attempting to process data that is new to them, which may or may not be defined, and may be messy or offering little value (Franks 2012).

In particular the act of integrating data from multiple sources with data formats of different types, i.e. tables, blog entries, tweets, emails, articles, where often mapping between schemas and linking records becomes problematic due to the heterogeneous nature, the absence of or incompleteness of schemas and the sheer volume of data (McAfee et al. 2012; Dong and Srivastava 2013) presents the most challenge for entity disambiguation and record linking. These dimensions, encompassed by volume, shown in Figure 1, contribute to what can be described as wild data (Lohr 2012) or data that is hard to manage due to its perplexity.

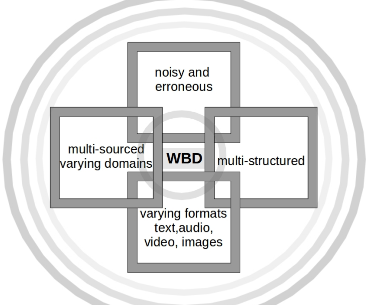
noisy and erroneous
multi-sourced varying domains
WBD
multi-structured
varying formats text,audio, video, images

Figure 1: Wild Big Data

Wild Big Data

Since data can be structured in multiple ways (unstructured, structured, semi-structured, and multi-structured (Franks 2012)), how we extract information becomes a challenge. The tools and algorithms we use for structured data are quite different than what we would use for unstructured data. This is even more challenged by the fact that there can be a mix within a single document. For example, unstructured text can include embedded structure such as tables. The same is true in how we process text vs. video or audio. We typically use different approaches given the format of the data.

In addition, data coming from multiple sources may be semantically represented in different ways. Sources can be related to different domains or the same domain but representations can be quite different. This has great impact on the accuracy of record linking and entity disambiguation.

However, what is often not emphasized is the fact that this multi-sourced data quite frequently can be schema-less. Data exported from a social networking web site, data coming from various sensors, for example may not have supporting or accessible schemas. Processing data is a challenge under these conditions and new approaches need to be established to deal with these demands. A company not prepared to process big data may incur loss if their competitors are prepared to do so (Franks 2012). Noise and erroneous data further complicates this task as they affect accuracy of methods used that try to make sense of new data which is not defined.

Processing Schema-less Data

Our work focuses on how to process data that is schema-less in order to perform record linking or entity disambiguation. We believe this is one important step in taming wild data. Our approach, which is common in the database domain and still relevant for Big Data, is to map unknown entities to known classifications of entities.

Fine-Grained Entity Type Identification

By identifying entities and their fine-grained types, i.e. soccer player rather than person, we provide a first step in making sense of wild data in the absence of schemas and when data is heterogeneous such that entities can originate from various domains identified by many types. We do this by means of mapping unknown types to known types defined by large open ontologies such as Freebase (Bollacker et al. 2008) and DBpedia (Auer et al. 2007). However, since these large ontologies can be composed of many domains and subdomains, we contextualize the process which can reduce the number of candidate types to a subset of types relevant to the domain or subdomains.

Our previous work outlined preliminary work that identified fine-grained entity types for heterogeneous graphs and performed experiments using DBpedia and Freebase (Sleeman, Finin, and Joshi to appear 2014). This work used information gain and a supervised approach to map unknown entities to known entities. We established what we described as 'high potential predicates' using information gain and mapped these to known entity types found in DBpedia. We then performed evaluations using Freebase and ArnetMiner (Tang, Zhang, and Yao 2007; Tang et al. 2008) data sets to determine how well we could identify entity types in Freebase and ArnetMiner by the entities described in DBpedia. We performed well (over 90% F-Measure on average for Person types) identifying ArnetMiner entity types since they were mostly *Person* types. However we did not perform as well with Freebase data since the Freebase types are broader than DBpedia types and there was less representation.

More importantly we found that there is an implicit context based on the set of unknown instances. If we could establish that context initially, we could reduce the number of candidate entity types. With large ontologies, the number of unique types can be large and eliminating types that are not relevant will both improve accuracy and reduce total computation time.

Table 1: Mapping Types

DBpedia Types	Freebase Types
Island	Island
Mountain	Mountain
MountainRange	MountainRange
NaturalPlace	?
River	River
Stream	?
Cave	Cave
LunarCrater	Lunar Crater
Valley	?
Volcano	Volcano
BodyOfWater	BodyOfWater,BodyOfWaterExtra
RiverBodyOfWater	BodyOfWater,BodyOfWaterExtra
?	US National Parks
?	Geographical Feature
?	Waterfall
?	Lake
?	Glacier
?	Rock Type

In this paper we introduce our preliminary work that contextualizes this mapping by the use of topic models.

Motivation

As in our previous work, our goal is to map unknown entity types to known entity types. We do this as precursor step for entity disambiguation. This process allows us to create coarse clusters of entities of the same fine-grained type, reducing the number of evaluations that would need to be performed for determining which entities are the same or similar. The key contribution is the identification of fine-grained types of unknown entities. This is important for WBD because entities are heterogeneous, obtained from different sources with potentially different representations. Entities can be extracted and combined from different source domains. This level of complexity in conjunction with volume is a challenge for entity disambiguation algorithms.

Mapping Types

Part of the mapping problem is that representations can be at different granularities and one-to-one mappings may not be feasible. For example, in Table 1 we show one way to map between DBpedia types and Freebase types for context *natural places*. There is not always a clear one-to-one mapping hence if we used DBpedia to train a model, we would have insufficient coverage for almost 40% of Freebase *natural places* types.

Background

Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is a well known generative graphical model that models latent topics and has been used to solve a number of classification problems including spam detection, author classification, short text and tweets (Bíró, Szabó, and Benczúr 2008; Phan, Nguyen, and Horiguchi 2008; Ritter et al. 2011;

Yan et al. 2013). In LDA a document is viewed as a multinomial distribution of topics β and a topic is viewed as a multinomial distribution of words Φ. Inference cannot be exact (Blei, Ng, and Jordan 2003), therefore estimation methods are used. There are a number of methods that could be used to perform the parameter estimation, however a standard approach is to use collapsed Gibb sampling, a Markov chain Monte Carlo algorithm (Griffiths 2002).

Approach

We use Freebase as our 'known' ontology and create a topic model using JGibbLDA (Phan and Nguyen 2006). The Freebase data F is in the form of triples where a triple t is composed of a subject node s, a predicate relationship p, and object node o. Defined by the following: $s \in (URI \cup Blank)$,$p \in (URI)$ and $o \in (URI \cup Blank \cup Literal)$ (Beckett 2004; Brickley and Guha 2004). An entity instance is defined by a set of triples T containing a common s URI.

Two-Step Training

For each $s_1 ... s_n \in F$, where n is the total number of subjects, we tokenize $t_1 ... t_m \in T$, where m is the total number of triples for a s, eliminating stop words and other commonly occurring words across entities. Where T represents a single document $d \in D$, which is used to train the model. We parameterize the number of topics but for our initial work we used 500 topics. Ongoing work will include experiments that measure performance based on the number of topics.

The second step in training is to map each known Freebase entity type ft to $\vec{topic}$ by means of inference. A document d is a set of triples from for all entity instances for a particular ft. This provides the basic mapping from freebase types to topics.

Mapping from Known to Unknown

For unknown DBpedia instance data DB, for each $s_1 ... s_n \in DB$, we tokenize $t_1 ... t_m$ and run inference for each, where a d is the set of triples for a s. We then use the same method to obtain a $\vec{topic}$.

Using KullbackLeibler divergence (Kullback and Leibler 1951), for each $s_1 ... s_n \in DBDS$, for each $\vec{topic}_1 ... \vec{topic}_l$ we then calculate how similar $\vec{topic}$ is to each $\vec{topic}$ associated with each ft. Given a constraining parameter, we define the number of types we associate with each unknown entity instance with associated probabilities.

The result is clusters of entity instances given their fine-grained type associations. We could then apply a clustering method to perform entity disambiguation.

Preliminary Results

We took a random sample of Freebase data that included 1218 different types and 5000 distinct entities with an average of 3 entity types per entity. We also took a random sample of DBpedia data that included 120 different types and 150 different entities with an average of 4 different entity types per entity. We built our topic model using Freebase data with the number of topics = 800.

Our preliminary experiment looked at DBpedia entities having types within 3 context categories (*sports*, *natural places*, and *creative works*). We manually mapped Freebase types to DBpedia types for each context. *Sports* included any type sports related (i.e. players, teams, locations), *natural places* included types such as rivers and mountains, and *creative works* including types related to television, music and books. 587 entities related to *sports* context, 1459 entities related to *creative works*, and 99 entities related to *natural places*. We then measured how many predicted types fell into each context category.

This experiment shows promise for our more comprehensive experiments which will measure exact type to type predictions. Our results showed a bias for sports related types for DBpedia sports instances in Figure 4 and a bias for creative work types for DBpedia creative work instances in Figure 3. As seen in Figure 2 Natural places did not show a strong bias for natural places types but we believe this is due to the lack of coverage in our training samples.

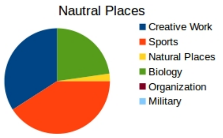

Figure 2: Natural Places DBpedia Instances

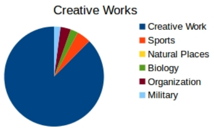

Figure 3: Creative Works DBpedia Instances

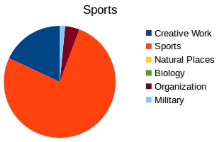

Figure 4: Sports DBpedia Instances

Interestingly sports players such as soccer players and football players, had a bias for sports types and a second bias for creative types, in particular television. Where as for sports stadiums, we saw a second bias for locations types rather than creative types. We saw similar interesting biases among creative types.

Related Work

Berlin et al. (Berlin and Motro 2002) performed database mappings using the "dictionary" approach that we suggest in our work. Their system which mapped database attributes to a common dictionary performed well producing over 70% harmonic mean. They however had domain experts manually annotate attribute mappings. This early work inspired our ideas for mapping to a known ontology.

In work by Biro et al. (Bíró, Szabó, and Benczúr 2008) they used a modified version of LDA to perform multi-corpus classification of spam. They ran LDA for each each class then created a union of the results of their topic collection. We find this work to be an interesting approach for classification and plan to consider this work for our future experimentation.

Most closely related to our work is research by Ritter et al. (Ritter et al. 2011) who uses a similar mapping approach to perform named entity type classification in tweets. Similar to our work, they map to Freebase types. However, their goals are slightly different, we use our type identification to help us identify entities. They use entity identification to help identify types. They used a modified LDA model for labeled data.

Paulheim et al. (Paulheim and Bizer 2013) perform entity type identification for DBpedia entities using a probabilistic method based on the instance data, using existing type identification as a inference point for identifying unknown types. Their work is largely dependent upon link analysis by which they generate their models they use for inference.

Ling et al. (Ling and Weld 2012) describe fine-grained entity recognition which uses fine-grained entity types to support entity recognition using an adapted perceptron algorithm. However they are addessing the problem of recognizing entity types for unstructured text.

Conclusions and On Going Work

We have formally described Wild Big Data, a natural result of the components of Big Data. We described an important aspect of this, schema-less data and its effect on entity disambiguation and record linking. We also presented preliminary results of our entity type mapping approach that incorporates a topic model to perform known type to unknown type mappings. Our ongoing work will include more experiments and larger data sets, specific to a big data domain.

References

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. Dbpedia: a nucleus for a web of open data. In *Proc. 6th Int. Semantic Web Conf.*, 722–735. Berlin, Heidelberg: Springer-Verlag.

Beckett, D. 2004. Rdf/xml syntax specification. http://www.w3.org/TR/REC-rdf-syntax/.

Berlin, J., and Motro, A. 2002. Database schema matching using machine learning with feature selection. In *Proc. Conf. on Advanced Information Systems Engineering*, 452–466. Springer.

Bíró, I.; Szabó, J.; and Benczúr, A. A. 2008. Latent dirichlet allocation in web spam filtering. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, 29–32. ACM.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proc. ACM Int. Conf. on Management of Data*, 1247–1250. ACM.

Brickley, D., and Guha, R. 2004. Resource description framework (rdf) schema specification 1.0. http://www.w3.org/TR/rdf-schema/.

Chen, H.; Chiang, R. H.; and Storey, V. C. 2012. Business intelligence and analytics: From big data to big impact. *MIS quarterly* 36(4):1165–1188.

Dong, X. L., and Srivastava, D. 2013. Big data integration. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. IEEE.

Franks, B. 2012. *Taming the big data tidal wave: Finding Opportunities in Huge data streams with advanced Analytics*, volume 56. John Wiley & Sons.

Griffiths, T. 2002. Gibbs sampling in the generative model of latent dirichlet allocation.

Hendler, J. 2013. Broad data: Exploring the emerging web of data. *Big Data* 1(1):18–20.

Hitzler, P., and Janowicz, K. 2013. Linked data, big data, and the 4th paradigm. *Semantic Web* 4(3):233–235.

Kullback, S., and Leibler, R. A. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 79–86.

Ling, X., and Weld, D. S. 2012. Fine-grained entity recognition. In *AAAI*.

Lohr, S. 2012. The age of big data. *New York Times* 11.

Mayer-Schönberger, V., and Cukier, K. 2013. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McAfee, A.; Brynjolfsson, E.; Davenport, T. H.; Patil, D.; and Barton, D. 2012. Big data. *The management revolution. Harvard Bus Rev* 90(10):61–67.

Paulheim, H., and Bizer, C. 2013. Type inference on noisy rdf data. In *International Semantic Web Conference*.

Phan, X.-H., and Nguyen, C.-T. 2006. Jgibblda: A java implementation of latent dirichlet allocation (lda) using gibbs sampling for parameter estimation and inference.

Phan, X.-H.; Nguyen, L.-M.; and Horiguchi, S. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, 91–100. ACM.

Ritter, A.; Clark, S.; Etzioni, O.; et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1524–1534. Association for Computational Linguistics.

Sleeman, J.; Finin, T.; and Joshi, A. to appear, 2014. Entity type recognition for heterogeneous semantic graphs. In *AI Magazine*. AAAI Press.

Tang, J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L.; and Su, Z. 2008. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, 990–998.

Tang, J.; Zhang, D.; and Yao, L. 2007. Social network extraction of academic researchers. In *ICDM'07*, 292–301.

Wu, X.; Zhu, X.; Wu, G.-Q.; and Ding, W. 2014. Data mining with big data. *Knowledge and Data Engineering, IEEE Transactions on* 26(1):97–107.

Yan, X.; Guo, J.; Lan, Y.; and Cheng, X. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, 1445–1456. International World Wide Web Conferences Steering Committee.