Understanding a Large Corpus of Web Tables Through Matching with Knowledge Bases – An Empirical Study

Oktie Hassanzadeh, Michael J. Ward, Mariano Rodriguez-Muro, and Kavitha Srinivas

IBM T.J. Watson Research Center
Yorktown Heights, NY, USA
{hassanzadeh,MichaelJWard,mrodrig,ksrinivs}@us.ibm.com

Abstract. Extracting and analyzing the vast amount of structured tabular data available on the Web is a challenging task and has received a significant attention in the past few years. In this paper, we present the results of our analysis of the contents of a large corpus of over 90 million Web Tables through matching table contents with instances from a public cross-domain ontology such as DBpedia. The goal of this study is twofold. First, we examine how a large-scale matching of all table contents with a knowledge base can help us gain a better understanding of the corpus beyond what we gain from simple statistical measures such as distribution of table sizes and values. Second, we show how the results of our analysis are affected by the choice of the ontology and knowledge base. The ontologies studied include DBpedia Ontology, Schema.org, YAGO, Wikidata, and Freebase. Our results can provide a guideline for practitioners relying on these knowledge bases for data analysis.

Keywords: Web Tables, Annotation, Instance-Based Matching

1 Introduction

The World Wide Web contains a large amount of structured data embedded in HTML pages. A study by Cafarella et al. [6] over Google's index of English documents found an estimated 154 million high-quality relational tables. Subsequent studies show the value of web tables in various applications, ranging from table search [15] and enhancing Web search [1, 3] to data discovery in spreadsheet software [2, 3] to mining table contents to enhance open-domain information extraction [7]. A major challenge in applications relying on Web Tables is lack of metadata along with missing or ambiguous column headers. Therefore, a content-based analysis needs to be performed to understand the contents of the tables and their relevance in a particular application.

Recently, a large corpus of web tables has been made publicly available as a part of the Web Data Commons project [12]. As a part of the project documentation [13, 14], detailed statistics about the corpus is provided, such as distribution

of the number of columns and rows, headers, label values, and data types. In this paper, our goal is to perform a semantic analysis of the contents of the tables, to find similarly detailed statistics about the kind of entity types found in this corpus. We follow previous work on recovering semantics of web tables [15] and column concept determination [8] and perform our analysis through matching table contents with instances of large cross-domain knowledge bases.

Shortly after we started our study, it became apparent that the results of our analysis do not only reflect the contents of tables, but also the contents and ontology structure of the knowledge base used. For example, using our approach in tagging columns with entity types (RDF classes) in knowledge bases (details in Section 2), we observe a very different distribution of tags in the output based on the knowledge base used. Figure 1 shows a "word cloud" visualization of the most frequent entity types using four different ontologies. Using only DBpedia ontology classes, the most dominant types of entities seem to be related to people, places, and organizations. Using only YAGO classes, the most frequent types are similar to those from DBpedia ontology results, but with more detailed breakdown and additional types such as "Event" and "Organism" that do not appear in DBpedia results. Freebase results on the other hand are very different, and clearly show a large number of music and media related contents in Web tables. The figure looks completely different for Wikidata results, showing "chemical_compound" as a very frequent type, which is not observed in Freebase or YAGO types. This shows the important role the choice of knowledge base and ontology plays in semantic data analysis.

In the following section, we briefly describe the matching framework used for the results of our analysis. We then revise some of the basic statistics provided by authors of the source data documentation [14], and then provide a detailed analysis of the entity types found in the corpus using our matching framework. We end the paper with a discussion on the results and a few interesting directions for future work.

2 Matching Framework

In this section, we briefly describe the framework used for matching table contents with instances in public cross-domain knowledge bases. Although implementation of this framework required a significant amount of engineering work to make it scale, the methods used at the core of the framework are not new and have been explored in the past. In particular, our MapReduce-based overlap analysis is similar to the work of Deng et al. [8], and based on an extension of our previous work on large-scale instance-based matching of ontologies [9]. Here, we only provide the big picture to help understanding the results of our analysis described in the following sections.

Figure 2 shows the overall matching framework. As input, we have the whole corpus of Web Tables as structured CSV files on one hand and a set of RDF knowledge bases which we refer to as *reference knowledge* on the other hand. Based on our previous work on data virtualization [10], we turn both

(a) DBpedia Ontology Tags

(b) DBpedia YAGO Classes Tags

(c) Freebase Type Tags

(d) Wikidata Type Tags

Fig. 1. Word Cloud of Most Frequent Column Tags

the tabular data and RDF reference knowledge into a common format and store them as key-values on HDFS. For tabular data, the key is a unique URI identifying a column in an input table, and the values are the values that appear in the column. For reference knowledge input, the key is the RDF class URI, and the values are the labels of instances of that class. For example, URI `rep://webtables/23793831_0_4377639018067805567.csv/company+name` represents column with header `company+name` in file `23793831_0_4377639018067805567.csv` in the input data. The values associated with this URI are contents of the column, which in this case is a list of company names. An example of reference knowledge URI is `http://dbpedia.org/ontology/Company` which is the DBpedia ontology class representing entities of type "Company". The values associated with this URI are labels of instances of this type, which means a list of all company names in DBpedia.

The similarity analysis component of the framework takes in the key-values and returns as output a table with each record associating a column in an input table with a tag which is an RDF class in reference knowledge, along with a confidence score. This tag indicates a similarity between values associated with the column and the class in input key-values, based on a similarity measure. Our system includes a large number of similarity functions but for the purpose of this study, we focus on one similarity measure that is very simple yet accurate and powerful for annotation of tables. Similar to Deng et al. [8], we refer to this

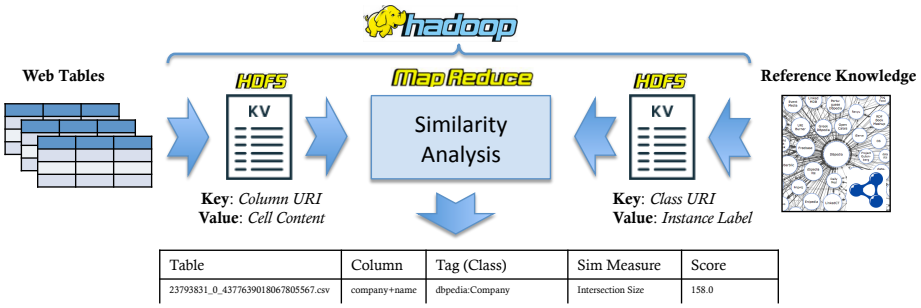


Fig. 2. Matching Framework

similarity analysis as *overlap analysis*. The values are first *normalized*, i.e., values are changed to lowercase and special characters are removed. We also filter numeric and date values to focus only on string-valued contents that are useful for semantic annotation. The similarity score is then the size of the intersection of the sets of filtered normalized values associated with the input URIs. The goal of overlap analysis is to find the number of values in a given column that represent a given entity type (class) in the input reference knowledge. In the above example, the column is tagged with class `http://dbpedia.org/ontology/Company` with score `158`, which indicates there are 158 values in the column that (after normalization) appear as labels of entities of type Company on DBpedia.

The *reference knowledge* in this study consists of three knowledge bases: (i) DBpedia [4] (ii) Freebase [5], and (iii) Wikidata [11, 16]. We have downloaded the latest versions of these sources (as of April 2015) as RDF NTriples dumps. DBpedia uses several vocabularies of entity types including DBpedia Ontology, Schema.org, and YAGO. We report the results of our analysis separately for these three type systems, which results in 5 different results for each analysis. We only process the English portion of the knowledge bases and drop non-English labels.

3 Basic Statistics

We first report some basic statistics from the Web Tables corpus we analyzed. Note that for this study, our input is the English subset of the Web Tables corpus [14] the same way we only keep the English portion of the reference knowledge. Some of the statistics we report can be found on the data publisher's documentation [14] as well, but there is a small difference between the numbers that could be due to different mechanisms used for processing the data. For example, we had to drop a number of files due to parsing errors or decompression failures, but that could be a results of the difference between the libraries used.

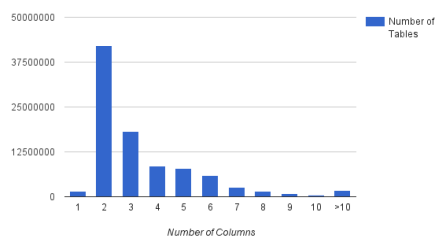
The number of tables we successfully processed is 91,357,232, that results in overall 320,327,999 columns (on average 3.5 columns per table). This results in 320,327,999 unique keys and 3,194,624,478 values (roughly 10 values per column) in the key-value input of Web Tables after filtering numerical and non-string

values for similarity analysis. DBpedia contains 369,153 classes, out of which 445 are from DBpedia Ontology, 43 are from Schema.org, and 368,447 are from YAGO. Freebase contains 15,576 classes, while Wikidata contains 10,250 classes. The number of values after filtering numeric and non-string values is 67,390,185 in DBpedia, 169,783,412 in Freebase, and Wikidata has 2,349,915 values. These numbers already show how different the knowledge bases are in terms of types and values.

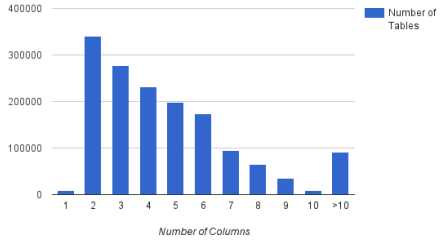
We first examine the distribution of rows and columns. Figure 3(a) shows the overall distribution of columns in the Web Tables. As it can be seen, the majority of the tables have lower than 3 columns. There are 1,574,872 tables with only 1 column, and roughly 62 million out of the 91 million tables (32%) have 2 or 3 columns. Now let us consider only the tables that appear in the output of our overlap analysis with intersection threshold set to 20, i.e., tables that in at least one of their columns have more than 20 normalized values shared with one of the knowledge reference sources. Such tables are much more likely to be of a higher quality and useful for further analysis and applications. Figure 3(b) shows the distribution of columns over these tables. As the figure shows, there is a smaller percentage of tables with small number of columns, with roughly 59% of the tables having 4 or more columns. This confirms the intuition that higher quality tables are more likely to have more number of columns, although there is still a significant number of tables with meaningful contents that have 3 or less columns.

Figure 3(c) shows the overall distribution of the number of rows in the whole corpus. Again, the majority of the tables are smaller ones, with roughly 78 million tables having under 20 rows, and roughly 1.5 million tables containing over 100 rows. Figure 3(d) shows the same statistics for tables with an overlap score over 20. Here again, the distribution of rows is clearly different from the whole corpus, with the majority of the tables having over 100 rows.

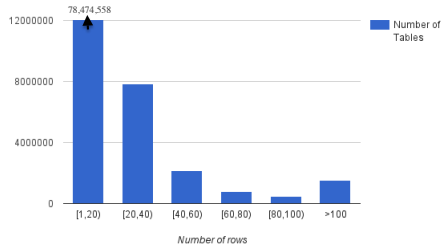
Next, we study the distribution of overlap scores over all tables and across different ontologies. Figure 4 shows the results (Schema.org results omitted for brevity). In all cases, the majority of tags have a score under 40, but there is a notable percentage of tags with a score above 100, i.e., the column has over 100 values shared with the set of labels of at least one type in the reference knowledge, a clear indication that the table is describing entities of that type. The main difference in the results across different ontologies is in the overall number of tags. With overlap score threshold of 20, there are 1,736,531 DBpedia Ontology tags, 542,178 Schema.org, 6,319,559 YAGO, 26,620,967 Freebase, and 865,718 Wikidata tags. The number of tags is a function of the size of the ontology in terms of number of classes and instances, but also the type system in the ontology. For example, Schema.org has only 43 classes resulting in an average of over 12,600 columns per each tag, but YAGO contains 368,447 classes which means an average of 17 columns per tag.



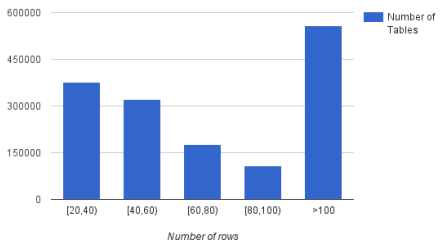
(a) Distribution of Number of Columns per Table



(b) Distribution of Number of Columns for Tables with 20+ Overlap Score



(c) Distribution of Number of Rows per Table



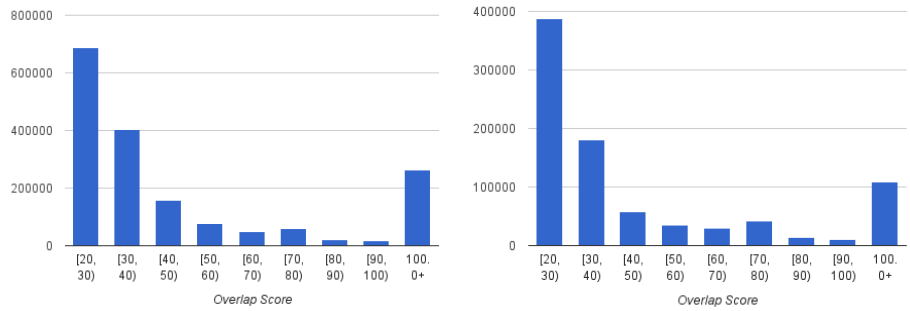
(d) Distribution of Number of Rows per Table with 20+ Overlap Score

Fig. 3. Distribution of Number of Rows and Columns

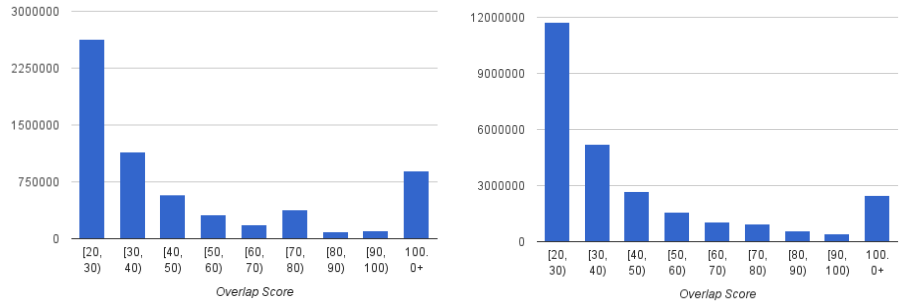
4 Distribution of Entity Types

We now present detailed statistics on the tags returned by the overlap similarity analysis described in Section 2. Going back to Figure 1 in Section 1, the word cloud figures are generated using the overlap analysis with the overlap threshold set to 20. The figure is then made using the top 150 most frequent tags in the output of the overlap analysis, with the size of each tag reflecting the number of columns annotated with that tag. The labels are derived either from the last portion of the class URI (for DBpedia and Freebase), or by looking up English class labels (for Wikidata). For example, "Person" in Figure 1(a) represents class `http://dbpedia.org/ontology/Person` whereas `music.recording` in Figure 1(c) represents `http://rdf.freebase.com/ns/music.recording`, and `chemical_compound` in Figure 1(d) represents `https://www.wikidata.org/wiki/Q11173` which has "chemical compound" as its English label.

In addition to the word cloud figures, Tables 1 and 2 show the top 20 most frequent tags in the output of our similarity analysis for each of the ontologies, along with their frequency in the output. From these results, it is clear that no single ontology on its own can provide the full picture of the types of entities that can be found on the Web tables. DBpedia ontology seem to have a better



(a) DBpedia Ontology (d) Wikidata



(a) YAGO (b) Freebase

Fig. 4. Distribution of Overlap Scores in Different Ontologies

coverage for person and place related entities, whereas YAGO has a large number of abstract classes being most frequent in the output. Schema.org provides a cleaner view over the small number of types it contains. Wikidata has a few surprising types on the top list, such as "commune of France". This may be due to a bias on the source on the number of editors contributing to entities under certain topics. Freebase clearly has a better coverage for media-related types, and the abundance of tags in music and media domain shows both the fact that there is a large number of tables in the Web tables corpus containing music and entertainment related contents, and that Freebase has a good coverage in this domain.

Finally, we examine a sample set of entity types across knowledge bases and see how many times they appear as a column tag in the overlap analysis output. Table 3 shows the results. Note that we have picked popular entity types that can easily be mapped manually. For example, Person entity type is represented by class `http://dbpedia.org/ontology/Person` in DBpedia, `http://dbpedia.org/class/yago/Person` in YAGO, `http://schema.org/Person` in Schema.org and

Table 1. Most Frequent Tags in DBpedia Ontology, YAGO, and Schema.org

DBpedia Ontology		YAGO		Schema.org	
Type	**Freq.**	**Type**	**Freq.**	**Type**	**Freq.**
Agent	242,410	PhysicalEntity	364,830	Person	186,332
Person	186,332	Object	349,139	Place	120,361
Place	120,361	YagoLegalActorGeo	344,487	CreativeWork	53,959
PopulatedPlace	112,647	Whole	230,667	Organization	50,509
Athlete	85,427	YagoLegalActor	226,633	Country	37,221
Settlement	60,219	YagoPerm.LocatedEntity	198,304	MusicGroup	22,926
ChemicalSubstance	57,519	CausalAgent	186,789	EducationalOrg.	12,159
ChemicalCompound	57,227	LivingThing	182,570	City	10,743
Work	53,959	Organism	182,569	CollegeOrUniversity	10,598
Organisation	50,509	Person	175,501	Movie	10,243
OfficeHolder	40,198	Abstraction	145,407	SportsTeam	9,594
Politician	39,121	LivingPeople	136,955	MusicAlbum	4,786
Country	37,221	YagoGeoEntity	120,433	Book	2,103
BaseballPlayer	30,301	Location	109,739	School	1,181
MotorsportRacer	26,293	Region	106,200	MusicRecording	1,166
RacingDriver	25,135	District	95,294	Product	1,130
Congressman	24,143	AdministrativeDistrict	92,808	TelevisionStation	1,037
MusicalWork	17,881	Group	85,668	StadiumOrArena	918
NascarDriver	16,766	Contestant	60,177	AdministrativeArea	896
Senator	15,087	Player	56,373	RadioStation	815

`http://rdf.freebase.com/ns/people.person` in Freebase. The numbers show a notable difference between the number of times these classes appear as column tags, showing a different coverage of instances across the knowledge bases. Freebase has by far the largest number of tags in these sample types. Even for the three ontologies that have the same instance data from DBpedia, there is a difference between the number of times they are used as a tag, showing that for example there are instances in DBpedia that have type Person in DBpedia ontology and Schema.org but not YAGO, and surprisingly, there are instances of Country class type in YAGO that are not marked as Country in DBpedia ontology or Schema.org.

5 Conclusion & Future Directions

In this paper, we presented the results of our study on understanding a large corpus of web tables through matching with public cross-domain knowledge bases. We focused on only one mechanism for understanding the corpus of tables, namely, tagging columns with entity types (classes) in knowledge bases. We believe that our study with its strict focus can provide new insights into the use of public cross-domain knowledge bases for similar analytics tasks. Our results clearly show the difference in size and coverage of domains in public cross-domain knowledge bases, and how they can affect the results of a large-scale analysis. Our results also show several issues in the Web Data Commons Web Tables corpus, such as the relatively large number of tables that contain very little or no meaningful contents.

Our immediate next step includes expanding this study to include other similarity measures and large-scale instance matching techniques [9]. Another interesting direction for future work is studying the use of domain-specific knowledge

Table 2. Most Frequent Tags in Wikidata and Freebase

Wikidata		Freebase	
Type	**Freq.**	**Type**	**Freq.**
Wikimedia_category	146,024	music.release_track	968,121
human	93,544	music.recording	964,906
chemical_compound	52,380	music.single	950,099
sovereign_state	34,681	location.location	532,053
country	22,030	people.person	475,472
determinator_for_. . . _occurrence	13,354	location.dated_location	460,766
city	12,823	location.statistical_region	458,643
commune_of_France	10,459	tv.tv_series_episode	440,985
taxon	10,127	location.citytown	409,315
landlocked_country	8,899	music.artist	390,458
island_nation	7,439	fictional_universe.fictional_character	372,820
republic	7,431	film.film_character	344,755
university	4,083	music.album	314,494
town	3,467	music.release	306,857
American_football_club	3,207	media_common.creative_work	304,231
band	3,024	media_common.cataloged_instance	297,875
municipality_of_Spain	2,950	type.content	269,216
comune_of_Italy	2,531	common.image	269,213
basketball_team	2,041	book.written_work	248,902
municipality_of_Germany	1,923	book.book	235,165

Table 3. Sample Entity Types and Their Frequency in Overlap Analysis Tags

Type	**DBpedia Ontology**	**YAGO**	**Schema.org**	**Wikidata**	**Freebase**
Person	186,332	175,501	186,332	93,544	475,472
Company	12,066	11,770	–	1,831	68,710
Location	120,361	109,739	120,36	–	532,053
Country	37,221	39,338	37,221	22,030	39,316
Film	10,243	9,080	10,243	348	175,460

bases to study the coverage of a certain domain in the corpus of Web Tables. For example, biomedical ontologies can be used in matching to discover healthcare related structured data on the Web.

The results reported in this paper may change after the reference knowledge sources or the corpus of tables are updated. Therefore, our plan is to maintain a website containing our latest results, along with the output of our analysis that can be used to build various search and discovery applications over the Web Tables corpus[1].

References

1. Google Web Tables. `http://research.google.com/tables`. [Online; accessed 29-04-2015].
2. Microsoft Excel Power Query. `http://office.microsoft.com/powerbi`. [Online; accessed 29-04-2015].
3. S. Balakrishnan, A. Y. Halevy, B. Harb, H. Lee, J. Madhavan, A. Rostamizadeh, W. Shen, K. Wilder, F. Wu, and C. Yu. Applying WebTables in Practice. In *CIDR*, 2015.

[1] For latest results, refer to our project page: `http://purl.org/net/webtables`.

4. C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - A Crystallization Point for the Web of Data. *JWS*, 7(3):154–165, 2009.
5. K. D. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, pages 1247–1250, 2008.
6. M. J. Cafarella, A. Y. Halevy, D. Zhe Wang, E. Wu, and Y. Zhang. WebTables: Exploring the Power of Tables on the Web. *PVLDB*, 1(1):538–549, 2008.
7. B. B. Dalvi, W. W. Cohen, and J. Callan. WebSets: extracting sets of entities from the web using unsupervised information extraction. In *WSDM*, pages 243–252, 2012.
8. D. Deng, Y. Jiang, G. Li, J. Li, and C. Yu. Scalable Column Concept Determination for Web Tables Using Large Knowledge Bases. *PVLDB*, 6(13):1606–1617, 2013.
9. S. Duan, A. Fokoue, O. Hassanzadeh, A. Kementsietsidis, K. Srinivas, and M. J. Ward. Instance-Based Matching of Large Ontologies Using Locality-Sensitive Hashing. In *ISWC*, pages 49–64, 2012.
10. J. B. Ellis, A. Fokoue, O. Hassanzadeh, A. Kementsietsidis, K. Srinivas, and M. J. Ward. Exploring Big Data with Helix: Finding Needles in a Big Haystack. *SIGMOD Record*, 43(4):43–54, 2014.
11. F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, and D. Vrandecic. Introducing Wikidata to the Linked Data Web. In *ISWC*, pages 50–65, 2014.
12. H. Mühleisen and C. Bizer. Web Data Commons - Extracting Structured Data from Two Large Web Corpora. 2012.
13. P. Ristoski, O. Lehmberg, R. Meusel, C. Bizer, A. Diete, N. Heist, S. Krstanovic, and T. A. Knller. Web Data Commons - Web Tables. `http://webdatacommons.org/webtables`. [Online; accessed 29-04-2015].
14. P. Ristoski, O. Lehmberg, H. Paulheim, and C. Bizer. Web Data Commons - English Subset of the Web Tables Corpus. `http://webdatacommons.org/webtables/englishTables.html`. [Online; accessed 29-04-2015].
15. P. Venetis, A. Y. Halevy, J. Madhavan, M. Pasca, W. Shen, F. Wu, G. Miao, and C. Wu. Recovering Semantics of Tables on the Web. *PVLDB*, 4(9):528–538, 2011.
16. D. Vrandecic and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.