A Statistical Comparison of Current Knowledge Bases

Michael Färber
Institute AIFB
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
michael.faerber@kit.edu

Achim Rettinger
Institute AIFB
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
rettinger@kit.edu

ABSTRACT

In the last years, many knowledge bases have been developed and used in real-world applications. These include DBpedia, Wikidata, and YAGO which all cover general knowledge and therefore similar topics. In this poster, we present statistical measurements on these KBs. Our experiments reveal that despite that fact that these KBs cover the same domains to a considerable amount, they differ from each other significantly w.r.t. their graph-based structure and ontological aspects.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

Keywords

Knowledge Bases, Knowledge Graphs, Statistics, Metrics

1. INTRODUCTION

In the last years, several knowledge bases (KBs) have been developed and found their way into industrial applications. Although KBs have been used a lot, to the best of our knowledge, comparative studies on the statistical characteristics of KBs are very limited so far. This is in particular true for the KBs DBpedia, Wikidata, and YAGO. These KBs are freely available and do not cover a specific domain, but general knowledge in general. In this paper, we focus on these KBs and exhibit their particularities w.r.t. their structural and ontological conditions. Based on the fact that these KBs are – from a conceptual point of view – directed graphs consisting of RDF triples,[1] we come up with simple graph-based and RDF-based metrics such as in-degree, out-degree and a variety of other metrics. Given the results of these metrics, we can gain a better insight into the particularities of these current KBs and learn to what extent they differ from each other.

[1]See `http://www.w3.org/RDF/`.

Hence, our main contributions in this paper are:

- We calculate a variety of statistical measurements on the widely used KBs DBpedia, Wikidata, and YAGO.
- We give an analysis regarding these results.
- We make our framework for statistical analysis of KBs available for the public[2] so that other KBs can be easily integrated.

The remainder of this paper is organized as follows: First we give an overview of related work of semantic graph analysis. We then introduce the KBs which we selected for our analysis, and provide details regarding the current versions of the KBs. We then present the results of applying several graph-based and semantics-based metrics on the KB datasets in question. After discussing particularities of our analysis in Section 3, we conclude in Section 4.

2. COMPARISON OF KNOWLEDGE BASES

2.1 Related Work

Firstly, some work on the analyis of the graph structure of the (HTML) Web has been carried out. Early studies of Web topology were published already in the 1990s (see, e.g., [2]). In 2000, Broder et al. [3] found out that the structure of the Web can be modeled in the shape of a bow tie. Rather recently, Donato et al. [4] developed some models which were brought into accordance with their crawl dataset regarding some characteristics such as the power law distribution for degree.

Secondly, related work has been carried out on the analysis of the Linked Open Data (LOD) cloud:[3] Rodriguez [9], for instance, analyzed the graph of data sources in the LOD cloud. Among other things, he concluded that, despite the general assumption of the LOD cloud being a crowded "ravel", the LOD cloud can be disaggregated into a component around DBpedia and another component around DBLP.[4] Gueret et al. [6] confirmed that observation, but added a third component around UniProt.[5]

Thirdly, a few analyses of single ontologies [11, 7, 5] were made – as we do it in this paper: Theoharis et al. [11] focus on power-law degree distributions. According to them,

[2]The implementation of the framework is available for download at `http://www.aifb.kit.edu/web/KB-Statistics`.
[3]See `http://lod-cloud.net`.
[4]See `http://dblp.uni-trier.de`. DBLP contains bibliographical information and is not domain-independent.
[5]See `http.//www.uniprot.org`.

ontologies exhibit power law degree distributions as soon as they have a sufficient number of predicates or classes. In this paper, we also calculate degree distributions and examine whether they follow a power-law. Hoser et al. [7] applied social network analysis on the two ontologies SWRC[6] and Suggested Upper Merged Ontology (SUMO).[7] According to the authors, eigenvalue analysis provides deep insights into the structure and focus of the ontology. In our work, in contrary, we do not take eigenvectors into consideration. In the context of describing and evaluating a benchmark generator for Linked Data, Duan et al. [5] used measurements such as indegree and number of distinct subjects/objects of specific KBs such as DBpedia and YAGO (as of 2011). Their work is therefore mostly related to our work. Duan et al. found out that there is a bad fit between the degree distribution of the Semantic Web benchmark and curated Linked Data datasets. They propose a new metric called coherence since the existing graph-based metrics do not make a point about the quality of a KB. However, as we see in our experiments, this metric is not properly applicable for our KBs.

2.2 Overview of the Knowledge Bases

In the following, we shortly describe the different KBs which we analyze in the following sections. We focus on these three KBs since they cover general, cross-domain knowledge and similar topics.

- **DBpedia**: DBpedia[8] is the most popular and prominent KB in the LOD cloud [1]. Since the first public release in 2007, DBpedia is updated roughly once a year.[9] DBpedia is created from automatically-extracted structured information contained in the Wikipedia, such as from infobox tables, categorization information, geocoordinates, and external links. Due to its role as the hub of Linked Open Data, DBpedia contains many links to other datasets in the LOD cloud. DBpedia is used extensively in the Semantic Web research community, but is also relevant in commercial settings: companies use it to organize their content, such as the BBC [8] and the New York Times [10]. In our experiments, we use the latest version of DBpedia, which is DBpedia 2014.[10]
- **Wikidata**: Wikidata[11] started on October 30, 2012 as a project of Wikimedia Deutschland. The aim of the project is to provide data which can be used by any Wikimedia project, including Wikipedia. Wikidata does not only store facts, but also the corresponding sources, so that the validity of facts can be checked. Labels, aliases, and descriptions for entities in Wikidata are provided in more than 350 languages. Wikidata is a community effort, i.e., users collaboratively add and edit information. Also, the schema is maintained and extended based on community agreements. In the near future, Wikidata will grow due to the integration of Freebase data.[12] Our experiments on Wikidata are based on the Wikidata simple statements dataset from February 2015.[13]
- **YAGO**: YAGO[14] – Yet Another Great Ontology – has been developed at the Max Planck Institute for Computer Science in Saarbrücken since 2007. YAGO comprises information extracted from the Wikipedia, WordNet[15], and GeoNames.[16] As of March 24, 2015, YAGO3 is available, which we use in our experiments. Since the YAGO3 data set was not available in triple format at the time of the experiments, we transformed the available tsv files into the triple format.

2.3 Analysis of the Knowledge Bases

2.3.1 Number of Triples

Comparing the number of triples in the different KBs (see Figure 1a), we can see that YAGO has much more triples than DBpedia or Wikidata. One reason for that might be that in case of YAGO (and Wikidata) there was only one dataset with all covered languages given (containing labels in different languages), while for DBpedia we could restrict the KB to the English language. Wikidata is rather small, since knowledge stored in Wikidata was not extracted from one text corpus – as in case of DBpedia –, but created by users of the Wikidata community.

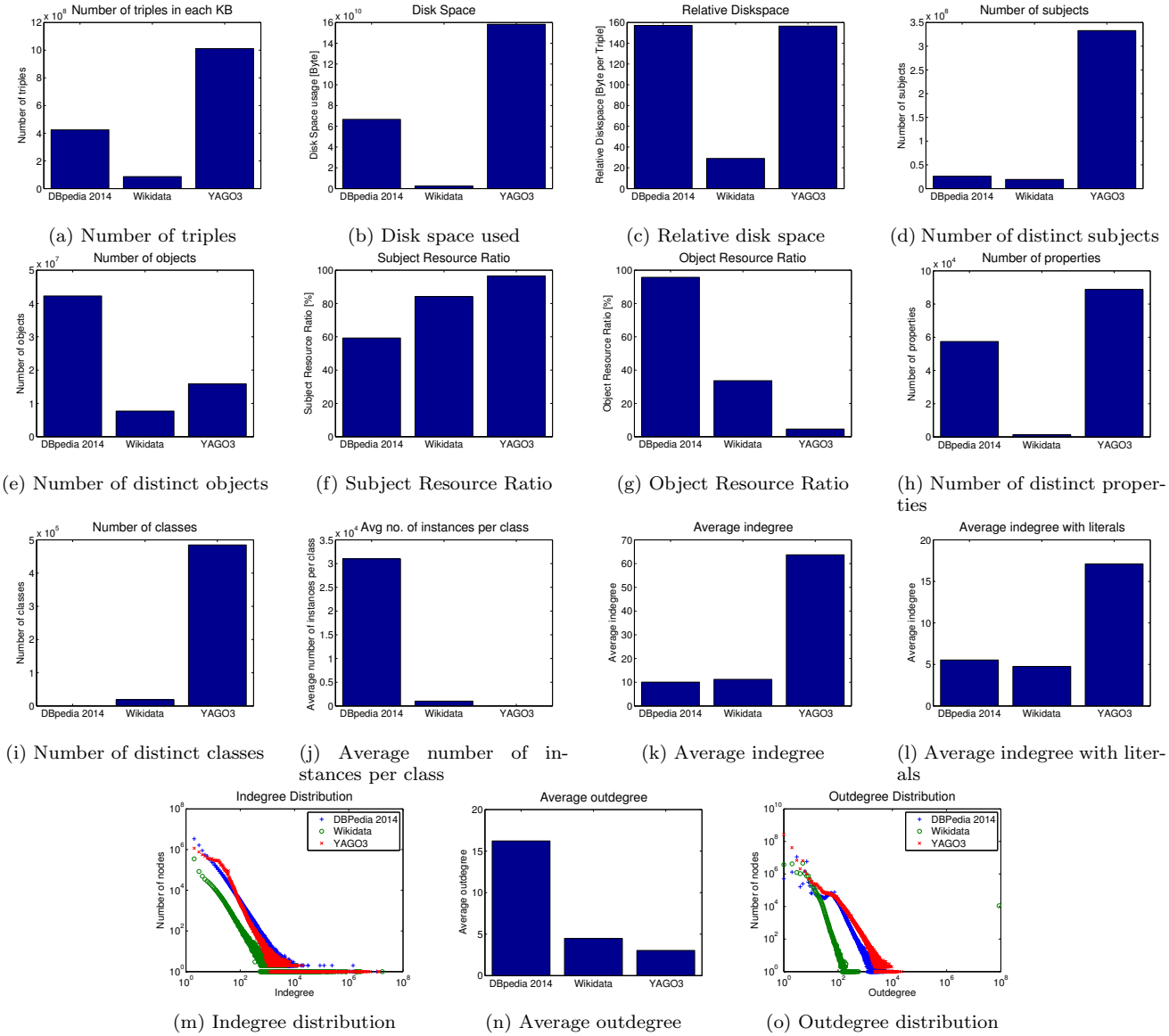
2.3.2 Disk Space

As visible in Figure 1b and as expectedly, the measured disk space is directly correlated to the number of triples. Figure 1c shows the relative disk space. Interesting is here the fact that – despite the relatively small number of triples – Wikidata requires much less disk space than the other KBs. The reason for that is that Wikidata uses non-human readable URIs (such as `http://wikidata.org/entity/Q1040`) while the other KBs rely on human-readable URIs (e.g., `http://dbpedia.org/resource/Karlsruhe` and `http://yago.org/resource/Karlsruhe`). In case of Wikidata, the human-readable labels for entities and properties are stored separately.

2.3.3 Number of Distinct Subjects and Number of Distinct Objects

Comparing the number of distinct subjects across the KBs in question (see Figure 1d) and the number of distinct objects (see Figure 1e), it becomes apparent that DBpedia has relatively few distinct subjects, but instead more distinct objects. In other words: The set of resources with outgoing edges is significantly smaller than the set of resources with incoming edges (ratio 1 : 1.6). YAGO, in contrast, has the opposite characteristic (ratio 21 : 1). Figure 1f and 1g show the ratio of the set of distinct subjects/objects w.r.t. to the entire set of resources in the KBs. Notable is that in case of YAGO, only to relatively few resources is linked.

[6]See `http://ontobroker.semanticweb.org/ontologies/swrc-onto-2001-12-11.oxml`.
[7]See `http://www.ontologyportal.org`.
[8]See `http://dbpedia.org`.
[9]There is also DBpedia live which is updated when Wikipedia is updated. See `http://live.dbpedia.org`.
[10]See our website for a list of the dump files used in our experiments.
[11]See `http://wikidata.org`.
[12]See `https://plus.google.com/u/0/109936836907132434202/posts/bu3z2wVqcQc`
[13]See `http://tools.wmflabs.org/wikidata-exports/rdf/exports/20150223/`.
[14]See `http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/`
[15]See `https://wordnet.princeton.edu`.
[16]See `www.geonames.org`.



(a) Number of triples (b) Disk space used (c) Relative disk space (d) Number of distinct subjects

(e) Number of distinct objects (f) Subject Resource Ratio (g) Object Resource Ratio (h) Number of distinct properties

(i) Number of distinct classes (j) Average number of instances per class (k) Average indegree (l) Average indegree with literals

(m) Indegree distribution (n) Average outdegree (o) Outdegree distribution

Figure 1: Statistics for the three KBs DBpedia, Wikidata, and YAGO.

2.3.4 *Number of Distinct Properties*

From our analysis regarding the number of distinct properties (see Figure 1h) we can derive that the used Wikidata RDF version contains only around 1,323 distinct properties. The reason for that is that properties are carefully introduced by the Wikidata community and go through an extensive discussion process before they are released for usage. DBpedia contains many properties. However, they are very heterogeneous and the non-mapping-based properties[17] (i.e., properties which were extracted not based on human-defined mappings, but solely as they appeared in the info-boxes in Wikipedia) are often very noisy.[18] A similar situation holds for YAGO.

[17]I.e. properties having the URI prefix `http://dbpedia.org/property/`.

[18]There are, for instance, 53,930 triples with the property `http://dbpedia.org/property/s` in DBpedia 2014 which has obviously no meaning.

2.3.5 *Number of Distinct Classes*

For calculating the number of distinct classes (see Figure 1i), we iterated over all instances contained in the KB datasets and took the objects of the relation `rdf:type`.[19] Although DBpedia often contains several classes according to this class-assignment method, we only retrieved 526 distinct classes. The small number in case of Wikidata can be justified again by the community approach of Wikidata. YAGO has a astonishing number of distinct classes since YAGO is mainly an ontology, i.e., containing class-based information such as the classes of the WordNet taxonomy. This last fact becomes apparent in Figure 1j where the average number of instances per class is visualized.

2.3.6 *Indegree*

Comparing the average indegree (defined as the average number of inlinks per node; see Figure 1k) where no triples with

[19]Standing for `http://www.w3.org/1999/02/22-rdf-syntax-ns#type`.

literals (values) on the object position were considered and comparing the average indegree where triples with literals were considered in addition (see Figure 1l), we can see that in general (i.e., for all KBs) the average indegree with literals is much lower than the average indegree where no literals were counted. The indegree for DBpedia and Wikidata is roughly the same. One reason might be that a considerable amount of Wikidata was taken from Wikipedia. It can be assumed that YAGO has a higher average indegree than DBpedia and Wikidata, since YAGO comprises many different ontologies.

The indegree distribution diagram (see Figure 1m) shows almost ideal logarithmic decreases of the number of nodes for all considered KBs. This is especially interesting since all KBs were created in different ways: automatically extracted from Wikipedia (DBpedia), partly created by the community (Wikidata), or composed of several sources which were used partly automatically, partly manually (YAGO). In the light of the figure we can also confirm that the power law is still applicable to the indegree distribution of semantic graphs such as the considered KBs.

2.3.7 Outdegree

Considering the average outdegree for each KB (defined as the average number of outlinks per node; see Figure 1n), we can see that nodes in the DBpedia knowledge graph have the highest number of outgoing links on average. Wikidata contains currently some domains of knowledge which are represented very densely (such as persons) while other domains are rarely covered yet. On average, however, Wikidata performs similarly as YAGO w.r.t. the average outdegree of nodes.[20]

The average outdegree of the KBs (see Figure 1o) suggest – as in the case of the average indegree – a power law distribution. However, if the outdegree is low, the power law distribution is broken. This confirms the theory of [11] which states that a sufficient number of predicates or classes is necessary for observing a power law distribution.

3. LESSONS LEARNED

According to Theoharis et al. [11], ontologies exhibit power law degree distributions as soon as they have a sufficient number of predicates or classes. Based on our experiments, we can confirm that for the KBs we considered.

Duan et al. [5] stated that "traditional" graph analysis metrics such as the degree or the number of classes are not suitable when KBs should be compared. Given our experimental results, we can confirm that to a certain extent. Duan et al. proposed a new metric called coherence metric where the "filling degree" of all entities of the different classes is calculated and aggregated. This might be a good indicator, however, the calculation for our KBs is tricky, since we often do not know the set of possible properties an entity of a specific class is able to have. Iterating over all existing properties of entities of this class is problematic since the KBs are often very noisy (different properties use the same meaning, different object types are used for the same property, etc.) and the considered KBs may contain multiple classes per instance.

[20]The outlier where the outdegree is 10^8 can be traced back to the fact that Wikidata contains many blank nodes with a high outdegree.

4. CONCLUSIONS

A measurement how current knowledge bases such as DBpedia, Wikidata, and YAGO look like and how they are structured, is to a large extent missing. In this paper, we presented a (freely available) framework for statistical analysis of KBs where any KB with triple format can easily be integrated. We calculated a variety of statistical measurements on the KBs DBpedia, Wikidata, and YAGO, since they all cover general knowledge and are used in many applications. Our investigations revealed that all current KBs performed very differently w.r.t. the presented metrics.

Acknowledgement

This work was carried out with the support of the German Federal Ministry of Education and Research (BMBF) within the Software Campus project SUITE (Grant 01IS12051).

5. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of the 6th ISWC and 2nd ASWC*, pages 722–735. Springer, 2007.

[2] T. Bray. Measuring the Web. In *Proceedings of the Fifth International World Wide Web Conference on Computer Networks and ISDN Systems*, pages 993–1005. Elsevier Science Publishers B. V., 1996.

[3] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000.

[4] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. The Web As a Graph: How Far We Are. *ACM Trans. Internet Technol.*, 7(1), Feb. 2007.

[5] S. Duan, A. Kementsietsidis, K. Srinivas, and O. Udrea. Apples and Oranges: A Comparison of RDF Benchmarks and Real RDF Datasets. In *Proceedings of the 2011 ACM SIGMOD*, pages 145–156, New York, NY, USA, 2011. ACM.

[6] C. Guéret, S. Wang, and S. Schlohbach. The Web of Data is a Complex System – First Insight into Its Multi-Scale Network Properties. In *Proceedings of the European Conference on Complex Systems*, pages 1–12, 2010.

[7] B. Hoser, A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Semantic Network Analysis of Ontologies. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, pages 514–529. Springer Berlin Heidelberg, 2006.

[8] G. Kobilarov et al. Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections. In *Proceedings of the 6th ESWC*, pages 723–737, Berlin, Heidelberg, 2009. Springer.

[9] M. A. Rodriguez. A graph analysis of the Linked Data cloud. *arXiv preprint arXiv:0903.0194*, 2009.

[10] E. Sandhaus. Semantic Technology at the New York Times: Lessons Learned and Future Directions. In *Proceedings of the 9th ISWC*, pages 355–355, Berlin, Heidelberg, 2010. Springer-Verlag.

[11] Y. Theoharis, Y. Tzitzikas, D. Kotzinos, and V. Christophides. On Graph Features of Semantic Web Schemas. *IEEE Trans. on Knowl. and Data Eng.*, 20(5):692–702, 2008.