Academic impact of DBpedia

Marcus Nitzschke

Abstract: There is an extensive collection of academic papers and presentations for the Wikipedia project available. This enables one to analyze the evolution of such projects over the years from the beginning. The goal of this paper is to create a similar collection for the DBpedia project. With this base there will be introduced several analyses and a comparison to the Wikipedia project.

1 Introduction

In 2011 there was a large survey to collect journal articles and conference paper concerning the Wikipedia project. The findings of this survey were published and partly analyzed[1]. Such an analysis showed that the number of conference paper decreased constantly from five years after the founding while the number of journal articles still grows.

A similar analysis for the DBpedia project is the goal of this paper. DBpedia extracts structured content from Wikipedia and republishes this content in a semantically understandable way. This leads to one of the most important datasets in the *Linked Open Data* cloud[2]. The DBpedia project was founded in 2007.

1.1 Important Conferences/Journals

The following listing introduces the abbreviations of the main conferences and main journals named in this paper.

- **ESWC** – *Extended Semantic Web Conference*
- **ISWC** – *International Semantic Web Conference*
- **LDOW** – *Linked Data on the Web*
- **WWW** – *World Wide Web Conference*
- **J. Web Sem.** – *Journal of Web Semantics*
- **SWJ** – *Semantic Web Journal*

[1] http://en.wikipedia.org/wiki/Wikipedia:Academic$_{\mathrm{studies\,of\,Wikipedia}}$
[2] http://richard.cyganiak.de/2007/10/lod/

2 Methods

This chapter describes which data sources were used and how the informations were retrieved and processed. This should lighten the process of replication for further examinations.

2.1 Data retrieval

The process of data retrieval included four sources. Table 1 gives an overview of these sources and the resulting number of publications.

Table 1: Overview of data sources

source	results
Google Scholar	84
Arnetminer	54
Semantic Web Conference Corpus	49
manually	4

The main criteria was that the term "dbpedia" occurs in the title or - if supported by the source - the abstract of the publication. Later the total amount was reduced to only match conference papers and journal articles which were all peer reviewed.

The following listing describes how the informations of the different data sources were retrieved.

- **Google Scholar**

  Google Scholar[3] allows to search a multiplicity of academic publications. The search was limited because the intent of this paper is to collect only the most relevant publications concerning DBpedia. The search term `allintitle:dbpedia` limited the total number of results to that containing "dbpedia" in their title. This search produced 84 results at the moment of writing (06/2012).

- **Arnetminer**

  Arnetminer[4] is a search engine which extends the collection of academic documents with informations about authors, conferences or journals. These informations linked together should enable a complete view of the data. In comparison to Google Scholar this also allows larger search possibilities. Besides the frontend search Arnetminer also provides a REST interface[5]. So it is possible to search the publications

[3] http://scholar.google.de/
[4] http://arnetminer.org/
[5] http://arnetminer.org/RESTful$_{service}$

by title. For better processing the retrieved data was converted from JSON to Bibtex by a small script[6].

- **Semantic Web Conference Corpus**

  The Semantic Web Conference Corpus[7] provides numerous informations about conferences and the presented talks belonging to the Semantic Web topic. This site also offers a SPARQL endpoint which was used to get the relevant publications for this paper. The query retrieved all publications with the term "dbpedia" in the title or in the abstract. A script[8] got the resulting URIs and scraped the corresponding html view for the Bibtex representation of the entry.

- **manually**

  The data sources listed so far could not cover all important journals like the *Semantic Web Journal* (SWJ). That's why a part of the results was collected manually. In case of the SWJ the underlying search term executed by a Google search was `dbpedia site:semantic-web-journal.net/content`. This means that all published research papers by the SWJ were searched by the term "dbpedia". The resulting pages were reviewed so that the term only occured in the title or abstract and not in a comment for example.

2.2 Data processing

To ensure a structured organization and analysis of the collected data a reference management software, namely Zotero[9], was used. This solution has numerous advantages, for example a Bibtex importer and an easy possibility to import results from Google Scholar via a browser plugin.

With the aid of Zotero the data processing first solved duplicates between the publications. This step reduced the total number of data entries from 191 to 130. Six further entries dropped out because of the type of publication, e.g. theses or technical reports, which are not in the scope of this paper. The second step was to complete lacking data fields like the publication date or conference names.

The cleaned data was exported from the Zotero database to the csv format by a SQL script[10]. This csv data is the base for the analyses which are done by Gnu R[11]. The main analyses provided in this paper are:

- total number of journal articles / conference paper by year
- total number of journal articles / conference paper by journal/conference

[6] TODO
[7] http://data.semanticweb.org
[8] TODO
[9] http://www.zotero.org/
[10] TODO
[11] http://www.r-project.org/

3 Results

The results contains the different analyses in the period between 2007 and 2012. Because this paper was written in June 2012 there is a lack of publications for this year. Especially conferences which are mostly held in the second half of the year are lacking, e.g. ISWC.

3.1 Analysis 1

The first analysis shown in Figure 1 plots the total number of publications by year. This is the analogical visualization of the Wikipedia analysis.

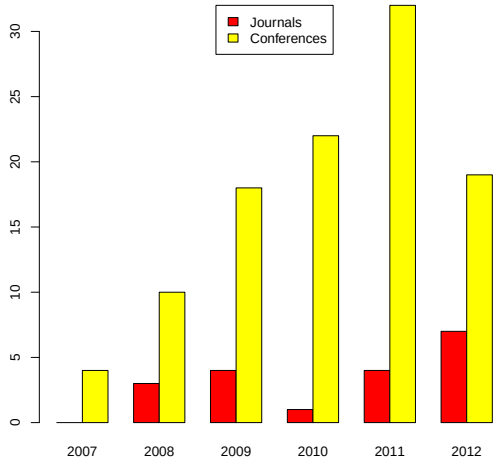

Figure 1: barplot visualizing total number of publications by year

The barplot shows a minimum number of publications in 2007 of zero and four for the journals and conferences respectively. The maximum number of journals is already reached in 2012 with seven. The maximum number of conference papers was reached in 2011 with 32. But it seems like this number will be exceeded in 2012.
There is an obvious trend that the number of conference papers is constantly growing. The trend of the journal articles is not that obvious because there is an outlier in 2010. But this fact is owed to the small number of journal articles in general. If we suppose that the number of journal articles is normal distributed the value of 2010 is still in the 95% confidence interval [1.19;5.14].
The reason why the number of conference papers is that greater than the number of journal articles was already covered by the Wikipedia analysis. “Journals are not the norm in CS/HCI research. Knowledge is shared through conferences, not journals.” [Kri11] they

explain. But there are opposite areas of research like the life sciences where journals have a higher relevance.

3.2 Analysis 2

The second analysis goes into the details of the journals and conferences. It plots the total number of publications by the specific journal or conference if there are at least two publications.

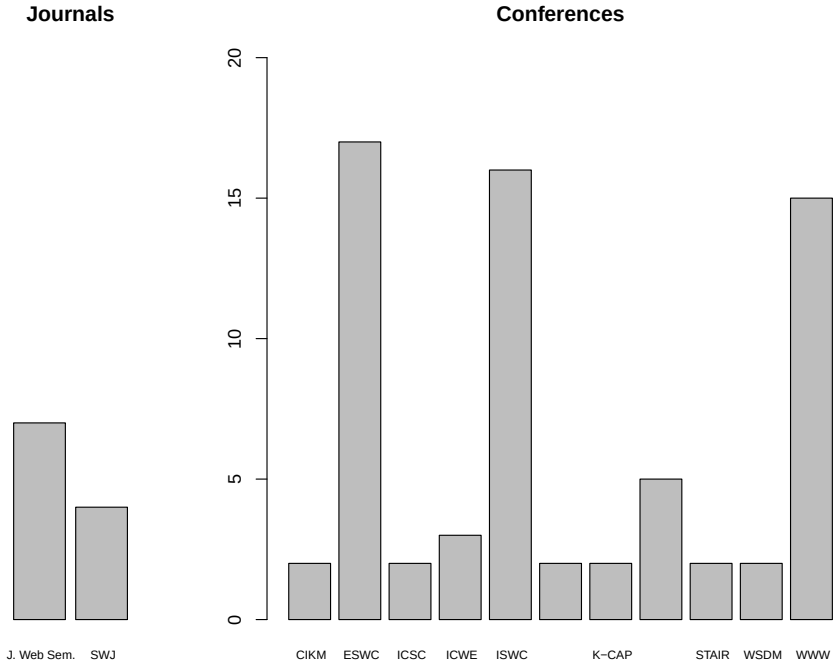

Figure 2: barplot visualizing total number of publications n by conference/journal with $n > 1$

Figure 2 illustrates that there are two journals which published more than one article related to DBpedia. The *Journal of Web Semantics* published seven articles, the *Semantic Web Journal* four articles. The second plot shows that there are three major conferences where DBpedia mattered since 2007. The most papers were published at *ESWC* with a number of 17. *ISWC* and *WWW* published 16 and 15 papers, respectively. It follows a gap of 10 to the *LDOW* conference with five published papers. This plot assumes the *ISWC/ASWC* as a seperate conference. Otherwise *ISWC* would have two additional publications and therefore 18 overall.

3.3 Analysis 3

Table 2 shows the importance of DBpedia at the major conferences in 2011. The table breaks the total number of publications at the conferences down to the number of publications where the term "dbpedia" occured somewhere in the paper and the number of publications where the term occured in the title or abstract.

Table 2: Compared number of publications at ESWC, ISWC, WWW in 2011
dbpedia_1 representing the occurence of "dbpedia" in full paper
dbpedia_2 representing the occurence of "dbpedia" in title or abstract

	ESWC (%)	ISWC (%)	WWW (%)
total	57 (100)	163 (100)	220 (100)
dbpedia_1	N/A	14 (8.59)	12 (5.45)
dbpedia_2	4 (7.02)	9 (5.52)	1 (0.45)

Therefore DBpedia was mentioned in 8.59% and 5.45% of the publications at ISWC and WWW somewhere in the paper. The highest rate of the dbpedia_2 criteria was reached by ESWC with 7.02% and a number of four papers. There were five more papers at ISWC, however this led to a lower percentage (5.52). The lowest percentage is reached by WWW with 0.45%.

3.4 Miscellaneous

Finally the collection of DBpedia related publications contains two highly cited papers. *"DBpedia: A Nucleus for a Web of Open Data"* [ABK+07] was cited 802 times based on Google Scholar. The second paper *"DBpedia - A Crystallization Point for the Web of Data"* [BLK+09] was cited 403 times and won the "JWS Most Cited Article 2006-2010 Award" in 2011.

4 Discussion

Compared to the Wikipedia study this paper showed mainly similar characteristics in the total number of publications per year analysis. The major difference is that the number of conference papers related to Wikipedia started to decrease six years after the foundation. This is exactly the year (2008) when DBpedia started to increase the number of conference publications. It will be interesting to see whether the importance of DBpedia will also start to decrease in the next one or two years.

One focus of this paper was to retrieve and process the data as transparent and reproducible as possible. That's why most of the publications were retrieved through several web services. Althought these services and the additional manual sources cover the bigger part and the most important publications it is hard to promise that all available publications that would observe the requirements of this study were retrieved. But these missing publications probably wouldn't lead to significantly different results.
The data sources also showed different measurements of quality. While Google Scholar contains the highest number of results it also contains the highest number of publications which had to be modified manually either because of lacking fields or the wrong type of publication. Besides Google Scholar this could also be caused by the Zotero plugin. The best quality of publications as well as possibility to retrieve the data was provided by the interface of the Semantic Web Conference Corpus. This showed the strengths of SPARQL and how it can improve searches.

References

[ABK+07] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, pages 722–735, 2007.

[BLK+09] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia-A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, 2009.

[Kri11] Travis Kriplean. Request to verify articles for Wikipedia literature review. http://lists.wikimedia.org/pipermail/wiki-research-l/2011-March/001366.html, 2011.