Linked Data and World Bank Open Data

Xunjie Liu
1613204
Department of Computer Science and Engineering
Xi'an Jiaotong-Liverpool University
Email: xunjie.liu16@student.xjtlu.edu.cn

***Abstract*—The purpose of the Semantic Web is to enhance the comprehension of machines towards data within web pages, so that enormous valuable could be fully utilized and reused. Linked Data was introduced aiming to realize the Semantic Web and applications based on Linked Open Data has been adopted in many areas. However, there is still great potential of Linked Data technology. In this article, the author present background of Linked Data and applications based on Linked Data. Background information, importance of Linked Data and current Linked Data application development are introduced in the literature review. Afterwards, the author will propose an idea of the combination of DBpedia and World Bank Open Data from two aspects, which are justification and feasibility.**

keywords: Semantic Web, Linked Data, DBpedia, World Bank Open Data

1. Literature Review

1.1. Introduction

Accompanying with the rapid development of Internet, information explosion brings both challenges and opportunities for traditional web structure. Although individuals and organizations publish numerous amount of data on web via protocols such as TCP/IP and HTML, very few of them could be understood by computers, since that the main content of traditional web technologies is designed for human to read rather than machines. On the current web the scale of content for computer to read and understand is very limited [1]. As a consequence, it is difficult to integrate and reuse data on the web if they are difficult to understand by computers. Besides, information explosion not only brings great amount of data, but also highly increases the complexity of data. With the requirement for more agile approaches to handle links and mash-up data in complex processes, conventional approaches of data integration would collapse in most situations [2]. In order to deal with these problems, the concept of Semantic Web is proposed and added to the current web layer to make the machine to understand web document.[3]

In order to support Semantic Web technology, the term Linked Data is introduced, referring to the method by which data could be described via Uniform Resource Identifier and Resource Description Framework, so that these data could be exchanged, displayed, linked and published [4]. In order to structure and standardize data so that it could be understood and utilized by machines, rules and standards are proposed such as Resource Description Framework (RDF) and Uniform Resource Identifier (URI). According to the rules raised by Berners-Lee in 2006, HTTP (Hypertext Transfer Protocol) URI strings is used to represent data so that people can locate data on internet wherever they are and different data could be linked together even if they are on the different servers, in this way data could be Linked. Then, standards such as RDF are utilized to structure data, to describe the relationship and other metadata, by means of these methods data could be understood by machines. Differs from conventional web structure whose primary units are HTML (HyperText Markup Language) documents connected by hyperlinks, Linked Data depends on documents containing data in RDF format [5]. Once data is machine-readable, which means it is explicitly defined, the link between local data set and external data set could be established [5].

1.2. Applications

Although Linked Data is still relatively new compared to the history of web, the practises of Linked Data have extended Web in a global scale, connecting different data from diverse fields such as people, companies, books, scientific publications, films, music, television and radio programmes, genes, proteins, drugs and clinical trials, online communities, statistical and scientific data, and reviews [5]. One of the first Linked Data project which start to realize the web of Linked Data was the Linked Open Data Project [6]. Depedia is a project aiming to extract data from Wikipedia and turning these data into the form of Linked Data so that enormous amount of data in Wikipedia could be readable and available for machines. It is described as one of the most famous components of the decentralized Linked Data effort [7]. DBpedia offers

tunnels for different application to get access to the data in Wikipedia, and many applications of Linked Data are based on it. For example, the BBC Music Beta project integrates not only database on its own servers, but also the information contained on Wikipedia, and it accomplished this aim by capturing data from DBpedia and develop its own data schema with data. Also, millions of geographical locations worldwide are extracted from Wikipedia and geogame by DBpedia [6]; in medical area, current Linked Open Data and technologies are used to automate the process of biomedical ontology generation [7]; search engine which crawls data from web via traversing RDF links performs more efficiently over large amount of data than traditional one [8]. Examples of applications of Linked Data are now distribute widely in many areas such as smart cities, academic communities, social media and healthcare.

According to the description of Tom Heath and Christian Bizer [9], to develop an application using Linked Data technology should follow 6 steps: accessing the web of data, vocabulary of mapping, identity resolution, provenance tracking, data quality assessment and finally, using data in the application context. In short, these 6 steps could be concluded as two parts: extract raw data from the web and integrate data according to its own data schema. Under current state of web, not most of the publishers would publish their web pages with Linked Data, and it would be more efficient to utilize a dataset of Linked Data than integrate data from all over the Internet. Websites such as Wikipedia, government official website, contain enormous amount of unstructured data, which is highly valuable for academic purpose or commercial purpose.

2. Linked Data and World Bank Open Data

2.1. Introduction of World Bank Open Data

World Bank Open Data is a website of The World Bank which provides free and open access to global development data in many indicators such as agriculture, education, environment and health. Similar to Wikipedia, World Bank Open Data contains tremendous amount of data from various areas. However, its main data is numerical digits with labels while types of information on Wikipedia are more various. For example, indicator of Forest area contains data of forest area for every country and the whole world, and data could be export as XML, CSV or Excel format file. Although data on World Bank Open Data is of great importance in many fields such as academic research, financial analysis, the operation of export or visualization of data still requires human to complete, which means that computer itself is hard to understand the data behind these tables.

2.2. Justification

The main idea of this essay is to transfer data on World Bank Open Data into Linked Data and link them with corresponding data in DBpedia, so that more powerful dataset could be available for application development. As it is stated in literature review, DBpedia is the fundamental of so many applications of Linked Data technology, and the first step of these applications is to retrieve required data from DBpedia. Therefore, if data on World Bank Open Data could be extracted in RDF format and link them to DBpedia, not only the content of DBpedia will be tremendously enrich, but also applications which based on DBpedia has opportunities to enhance their functions and enrich content.

One important feature of World Bank Open Data is that it manly contains numerical data such as GDP values of countries in the world, and those numbers are authoritative since that they are the result of government statistical departments. In traditional search engine, when users try to search for numbers for a specific item such as China GDP in 2018, the search engine mostly just compares the keywords with context in HTML files and return bunch of different relative web pages rather accurate numbers. The reason for this result is that machine cannot understand the meaning of those numbers and if authoritative numerical data could be readable by computers, the search process would be efficient and accurate. The improvement of accuracy of this kind of problems could be helpful for academic purpose and numerical data related fields.

The concept of Link Data search engines is conceived since the birth of Linked Data and there are already some good examples of Linked Data search engine such as Sig.ma, Falcons and SWSE[9]. These Linked Data search engines provide key-word based search which is designed for human-computer interaction, same as current prevalent search engines such as Google and Bing, while the content returned by Linked Data search engines is much richer than traditional search engines. However, neither traditional nor Linked Data search engines provide powerful function for numerical data search, computer cannot understand such question: GDP values of China from 2010 to 2018, since that DBpedia does not contain enough numerical data for computer to understand and select accordingly.

2.3. Feasibility

To develop a Linked Data search engine with the ability to handle numerical requests, steps of developing a Linked Data application mentioned in literature review could be followed.

2.3.1. Accessing the Web of Data. Before this application could access the structured data on DBpedia, data on

World Bank Open Data should be extracted firstly and then integrated into DBpedia. This process could be done easily since that data on World Bank Open Data could be exported as XML format and RDF documents are written in XML.

2.3.2. Vocabulary Mapping. Indicators or country names on World Bank Open Data should names after its URI in RDF and then they could be mapped to corresponding records in DBpedia. Once RDF links of World Bank Open Data and DBpedia are linked together, DBpedia is enriched and following steps (Identity Resolution, Provenance Tracking and Data Quality Assessment) are no need.

2.3.3. Using the Data in the Application Context. Differs from other search engines, this Linked Data search engine should firstly cache numerical data from DBpedia according to the keywords in request, then base on the predicates in request, do manipulation to cached data, finally return the result.

Nov. 28, 2018

3. Conclusion

Increasingly web publishers upload their web pages with Linked Data and applications based on Linked data technology would be popular in the future. Adoptions of maturing technology are supporting the deployment of Semantic Web, both in commercial and public organization (Motivation for the semantic web).

References

[1] Antoniou, G. , and Frank V. H. *A semantic web primer*. MIT press, 2004.

[2] S. Mouzakitis, D. Papaspyros, M. Petychakis, S. Koussouris, A. Zafeiropoulos, E. Fotopoulou, L. Farid, F. Orlandi, J. Attard and J. Psarras, "Challenges and opportunities in renovating public sector information by enabling linked data and analytics", *Information Systems Frontiers*, vol. 19, no. 2, pp. 321-336, 2016.

[3] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. Scientific American, 284(5), 2837. doi:10.1038/scientificamerican0501-34 PMID:11341160

[4] A. Hilario, "From Bibliographic Records to Data Changes in the Library Environment with the Application of Linked Open Data Technologies", *Information Resources Management Journal*, vol. 27, no. 3, pp. 28-41, 2014.

[5] C. Bizer, T. Heath and T. Berners-Lee, "Linked Data - The Story So Far", *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1-22, 2009.

[6] F. Bauer and M. Kaltenbock, *Linked open data: the essentials*. Wien: Ed. mono/monochrom, 2012.

[7] "Transcript: Sir Tim Berners-Lee Talks with Talis about the Semantic Web", *Web.archive.org*, 2018. [Online]. Available:https://web.archive.org/web/20130510134842/http:/talis-podcasts.s3.amazonaws.com/twt20080207_TimBL.html. [Accessed: 28- Nov- 2018].

[8] M. Alobaidi, K. Malik and S. Sabra, "Linked open data-based framework for automatic biomedical ontology generation", *BMC Bioinformatics*, vol. 19, no. 1, 2018.

[9] T. Heath and C. Bizer, "Linked Data: Evolving the Web into a Global Data Space", *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 1, no. 1, pp. 1-136, 2011.