Towards a Semantic Discovery for Heterogenous Open Data by Interlinking Metadata of Datasets

Jiseong Son[1], Youngsung Son[2], and Haklae Kim[1]

[1] Korea Institute of Science and Technology Information, Korea
{jsson,haklaekim}@kisti.re.kr
http://www.kisti.re.kr
[2] Electronics and Telecommunications Research Institute
{ysson}@etri.re.kr

Abstract. Open data refers to data that everyone can freely use, reuse and redistribute. A number of open data is released by various organisations, governments or communities. However, it is limited to discover datasets that users want, since most of data portals allow to search their datasets based on simple keywords using file names or descriptions, etc. This paper proposes a novel way for discovering disclosed government datasets by using linked data technologies. For achieving this objective, a set of datasets is collected from the public data portal in Korea, and all of data fields are extracted and transformed into linked data using an ontology model. We also provide a simple evaluation, which compares a search performance between the portal and the proposed method.

Keywords: Open Data · Government Data · Semantic Discovery · Ontology.

1 Introduction

While the big data phenomenon is becoming increasingly common, it is not easy for anyone to freely use the data. A large amount of big data is owned by service providers or platform owners, and only a limited portion of data is shared. On the other hand, open data allows users to provide a significant opportunity that they are able to use a variety of data across heterogeneous data sources and domains. The key value of open data is that a piece of data contained in published data can be interlinked with other data. In an open data environment, data can be interchanged between institutions, between institutions and governments, or between governments, and new value can be created through interlinking of datasets [2].

One of issues aligning on open data is that discovering datasets is getting difficult [1]. Most data portals provide the ability to discover datasets. For example, CKAN (Comprehensive Knowledge Archive Network), which is a data portal platform, is able to retrieve a file name and a description for files, and tags and file types added to the dataset [5, 6]. However, there is a limitation to searching for the information that an individual dataset has. If a user wants

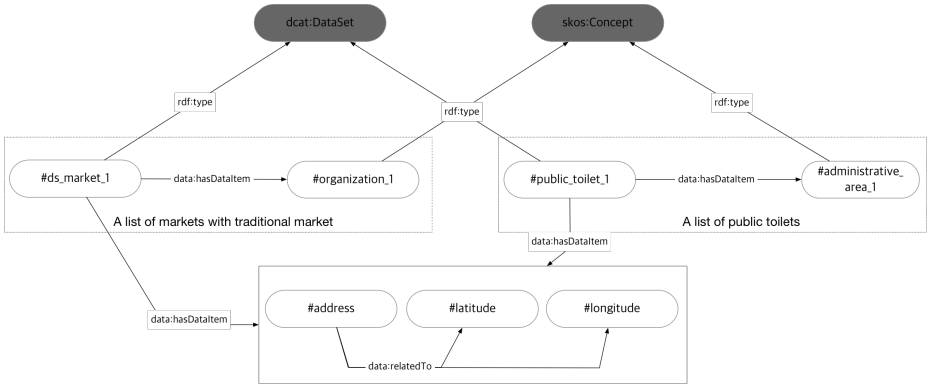

Fig. 1. A data model of representing a dataset and its metadata

to find out datasets that have 'population', most of data portals returns a list of datasets that contains the keyword (i.e. 'population') on descriptions or file names of the datasets on behalf of retrieving their content [4].

This study proposes a method of discovering disclosed government datasets by extracting data fields of individual datasets and constructing them as linked data. Section 2 describes a research approach including data collections and transformations based on a proposed ontology model. Section 3 introduces a small evaluation to retrieve the collected datasets with some comparisons. Section 4 concludes and introduces future research.

2 Research approach

We collect a set of public open datasets from the public data portal[3] and extract all of data fields from the datasets. This site provides governmental open datasets released by the Republic of Korea. Currently, 689 organisations provide 22, 334 file data (CSV or other types), 2,547 open APIs, and 91 standard data.

This paper focuses on analysing the standard data, since metadata quality of other datasets is not good to our purposes [3]. Note that the standard data in the portal refers to a set of datasets by using the public data open standard guidelines of the government that defines an item name (data field) and its value for 93 domains. A total of 1,480 item names were extracted in the collected standard datasets, there are 903 item names that eliminate redundancy. The selected fields are no needs for further clustering, since a data field is already normalised by using standardised terms. Note that the collected datasets containing the road-name address and the land-number address are 53 and 44, respectively, and the latitude and longitude include 55 data sets. Latitude and longitude data fields are defined together in all datasets. There are 5 cases where latitude/longitude is included in the dataset in which the road-name address as an item name does

[3] http://data.go.kr

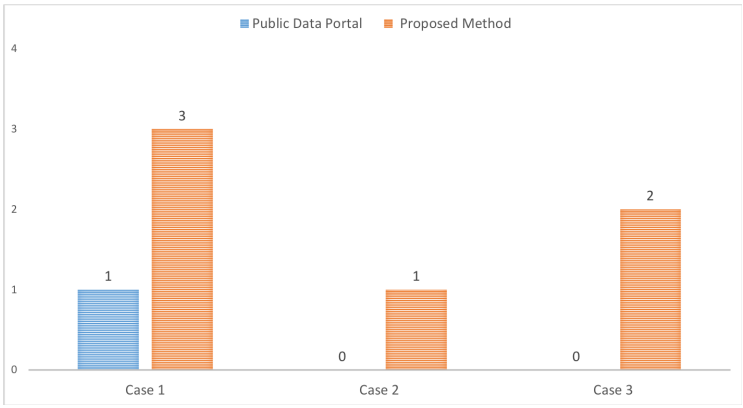

Fig. 2. The search results for the data portal and the proposed method

not exist, and there are 12 cases when there is no land-number address. On the other hand, when there is no latitude and longitude item name, the road-name address and the land-number address correspond to one case of three. There are 14 datasets that do not have both address and latitude/longitude information.

A simple ontology model is designed for representing a relationship between a dataset and its data field as shown in Figure 1. Each dataset has a set of data fields, and this relation is represented by using the data:hasDataItem property. Note that the data:relatedTo property is to describe a relationship between specific terms. For example, a 'location' may be related to 'address', 'latitude', or 'longitude'. There is no dataset with an item name of 'location', but most of datasets have 'address' or 'latitude and longitude'. In this reason, this property is used to expand a specific query. As shown in Figure 1, a traditional market dataset does not have any fields associated to a toilet. However, it is possible to discover some toilets around a traditional market, because both datasets have address or locational information.

3 Evaluation

We report the measurements obtained in Figure 2. We compare the three cases for the data portal and the proposed method. Case 1 discovers for a dataset with a single keyword. The portal and the proposed model have 1 and 3 results, respectively, for a specific topic (i.e. 'toilet'). Two of the results of the proposed model have no related keywords in the file name or description. Case 2 is a method for searching heterogeneous datasets. Consider the following query: what datasets contain a market and toilet information nationwide? Such queries are dependent on the information contained in the dataset. Although a particular dataset can be discovered if it has both fields, searching in a fragmented dataset is difficult. As shown in Figure 2, the portal does not have search results for multiple keywords (i.e. 'market' and 'toilet'), but the proposed model gives two

results. However, these results provide a simple information about a toilet as yes or no. Case 3 is to find out a specific relationship between datasets. For example, the data:relatedTo property can be used for discovering a relationship between a traditional market and a toilet. First, it retrieves a list of exact administrative area from both datasets based on address information, and then calculates a distance between search results using the latitude and longitude information. Compared to Case 2, this result show a specific location of a toilet around the market.

4 Conclusion

This paper proposes a new approach to discover datasets on a data portal by using linked data technologies. Most of data portals allow users to retrieve their datasets with search options, including keywords, data types, or user-generated tags, etc. However, it is limited to discover datasets based on their content. In this reason, users need to check whether these datasets are suitable to their purposes about search results. To solve this problem, this paper introduces a simple semantic search that aims to discover internal content of individual datasets by constructing linked data including data fields from individual datasets and its relationships. Although experimental data are relatively small, the evaluation shows that the proposed method is more effective than existing search methods. Future research will apply the data model and search method proposed in this paper to the whole data provided by the public data portal.

References

1. Hand, D.: Data, not dogma: Big data, open data, and the opportunities ahead. In: Tucker, A., Hppner, F., Siebes, A., Swift, S. (eds.) Advances in Intelligent Data Analysis XII, Lecture Notes in Computer Science, vol. 8207, pp. 1–12. Springer Berlin Heidelberg (2013)
2. Janssen, M., Charalabidis, Y., Zuiderwijk, A.: Benefits, adoption barriers and myths of open data and open government. IS Management **29**(4), 258–268 (2012)
3. Kim, H.: Quality evaluation of the open government data: The case of the open data portal of korea. International Journal of Contents (in press)
4. Kostovski, M., Jovanovik, M., Trajanov, D.: Open data portal based on semantic web technologies. In: Proceedings of the 7th Annual South-East European Doctoral Student Conference (DSC 2012). pp. 504–516 (2012)
5. Loureno, R.P.: Evidence of an open government data portal impact on the public sphere. IJEGR **12**(3), 21–36 (2016)
6. Scholz, R., Tcholtchev, N., Lmmel, P., Schieferdecker, I.: A ckan plugin for data harvesting to the hadoop distributed file system. In: Ferguson, D., Muoz, V.M., Cardoso, J.S., Helfert, M., Pahl, C. (eds.) CLOSER. pp. 19–28. SciTePress (2017)