Links to linked open data to increase the usefulness of data
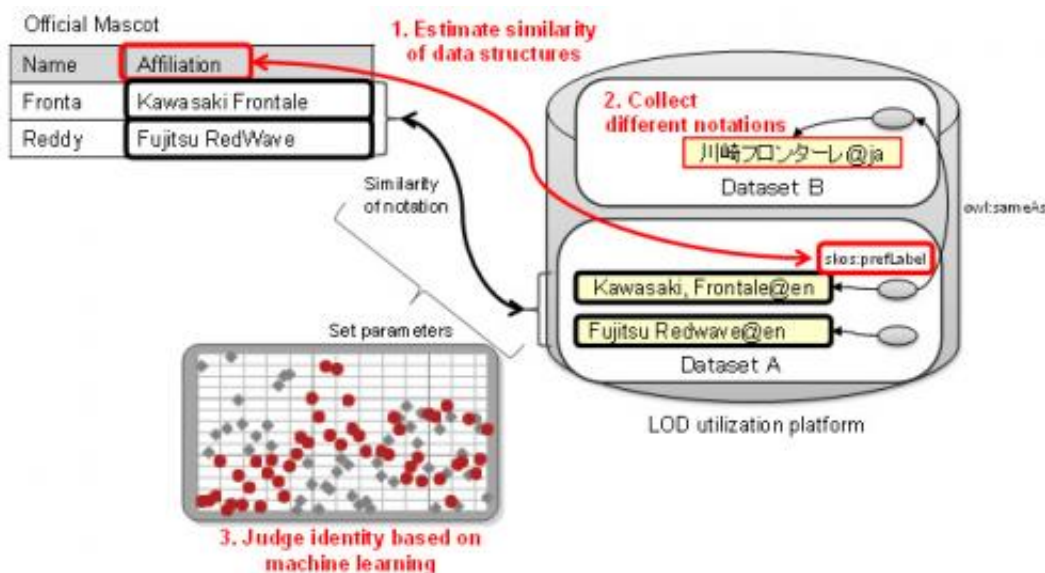
January 16 2014

Figure1. Overview of the new algorithm

Fujitsu Laboratories today announced the development of technology that can discover and automatically link data representing the same underlying subject among Linked Open Data (LOD) available throughout the world and individual data sets maintained by governments and companies.

LOD is starting to come into wider use as a mechanism for publishing data on the Internet. Each individual LOD record is intended to be linked to data published on other websites, and by following these links,

users can traverse multiple websites to access the data they need. When publishing data under the LOD approach, however, it can be challenging to interpret published data and determine which data is related in order to link to data on other websites.

The new technology enables inferences as to when data records refer to the same thing based on similarities in their notation and data structures, thereby making it possible to assign links. For example, the technology is expected to help increase the value of open data by making it possible to use LOD published by governments in combination with data held by companies and other LOD throughout the world.

In January, Fujitsu Laboratories is planning to launch a publically available search service for LOD data that makes it possible to tie in with the new technology: lod4all.net/

Background

Open data has rapidly garnered attention, as demonstrated by the release of the "Open Data Charter" at the G8 Summit in June 2013. In Japan, the IT Strategic Headquarters of the Japanese government's Cabinet has promulgated an e-gov open data strategy since July 2012, and declared the release of public data to the private sector (open data) to be one of the three pillars of the Cabinet's "Declaration of Creating the World's Most Advanced IT Nation" announced in June 2013.

In collaboration with the Irish Research Institute Insight Centre for Data Analytics, at National University of Ireland Galway (previously known as the Digital Enterprise Research Institute), Fujitsu Laboratories has developed an LOD utilization platform that can collect and perform batch searches on LOD published throughout the world.

Technological Issues

With LOD, it is advantageous that interrelated data, even data stored on different websites, be linked. This lets data users traverse multiple websites to access the data they need. However, when data is published on different websites, even if it represents the same underlying subject, differences in how it is structured or denoted cannot be resolved through simple keyword searches. As a result, data creators have been forced to find data they want to link to ahead of time, understand how that data is structured and denoted, and match it up to their own data.

In addition, because there had not been a means of traversing numerous websites to discover related data, data creators had been able to link only to data that they were already aware of. This means that while possible to link to well-known data sets and publish it in LOD format, it was difficult to link to data scattered across the web.

About the Technology

Fujitsu Laboratories has developed technology that leverages its LOD utilization platform to assign links based on similarities in notation and data structures. This makes it possible to automatically discover when multiple records refer to the same underlying subject. Features of the technology are as follows.

1. Technology for inferring when LOD data refers to the same person, organization, place, or other subject as that found in other data

Inferences are made by combining the following newly developed features:

- Resolving differences in data structures: Uses similarity in notation to measure the similarity of data structures.
- Resolving differences in notation: Uses the data structures in LOD to collect different notations about the same subject.
- Resolving ambiguity: Places parameters on similar data structures and notations and leverages machine learning to judge subject identity.

This technology achieved top-ranked inference accuracy in competitions in the US and China.

2. Ties in with LOD utilization platform

By tying in with the LOD utilization platform, which collects and performs batch searches on LOD published throughout the world, the technology can discover globally dispersed data that represents the same subject in different LOD datasets. So, for example, it can link to information not only in English-language data sets, but in other language data sets as well.
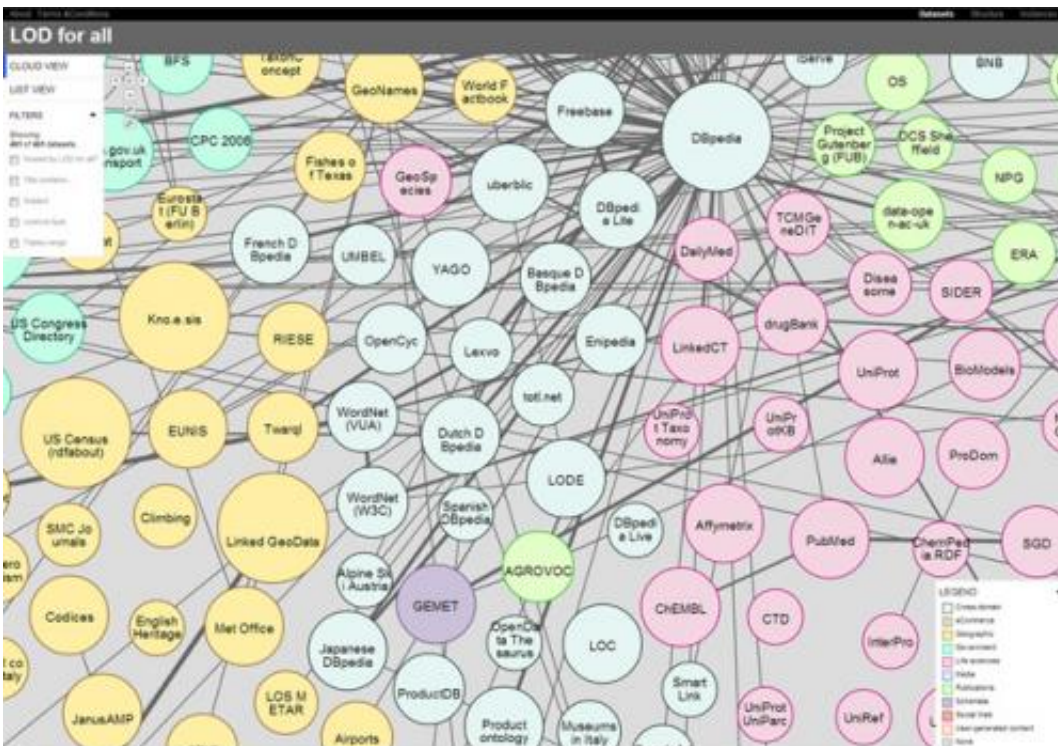
Figure2. Sample search interface display

Results

The newly developed technology makes it possible to discover and link data representing the same subject in multiple LOD datasets published around the world. This makes it simple to use a company's own data in combination with LOD data if, for instance, a national government publishes LOD data.

From January, Fujitsu Laboratories is planning to launch a LOD search service, available at lod4all.net/, that can tie in with the new technology. The search service features a visual, interactive search interface that takes advantage of the LOD utilization platform. From LOD datasets around the world that meet the service's license and download requirements, searches can be performed and the content of data viewed.

Future Plans

Fujitsu Laboratories is leveraging the newly developed LOD linking technology in a variety of field test projects with open data from national and local governments, with the aim of commercializing the technology in fiscal 2015.

Provided by Fujitsu