Web Data Integration: Student project abstract

Oliver Frendo, Dandan Li, Zehui Wang, Yi-Ru Cheng
October 4, 2015

1 DESCRIPTION OF USE CASE

The use case for this project is to integrate a dataset of companies listed at a certain stock exchange with multiple other datasets containing information about companies in general, in order to gather more information about companies than what is available from just one source. This will then be integrated with data about the cities their headquarters are located in. As such the resulting integrated dataset may be used for additional information regarding companies and their location around the world. We plan to use the following classes:

- Company
- Location

First we will gather data from each data source to create our datasets with the aim of enabling translation to the integrated schema. In most cases this means writing queries to different web services (DBPedia and Freebase) to request information about companies and locations. In one case the dataset is offered as .csv file and is ready to be used.

In the second phase the main challenge of identity resolution will be to accurately identify the companies that exist in multiple datasets. This will be done by using overlapping attributes that exist in multiple datasets.

Lastly conflicting information about companies, e.g. the location of their headquarters or the number of their employees, must be resolved by using specific resolution strategies per attribute. Finally the datasets can be merged and represented in the form of our integrated target schema.

2 DATASETS

2.1 SCHEMA AND BASIC PROFILE OF EACH DATASET

Dataset	Source	format	class	# entities	# attributes
	List of attributes				
London Stock Exchange	See (1)	csv	company	2.396	9
	List date, Company, Sector, Sub Sector, County of Incorporation, Market, Mkt Cap, International, Main Market, FCA Listing Category				
DBPedia	dbpedia.org/sparql	xml	company	11.966	9
	Founding year, name, keyPeople, industry, location, netIncome, operatingIncome, revenue, numberOfEmployees				
Freebase	freebase.com/query	json	company	3.142	7
	name, number_of_employees, revenue (currency), headquarters, leadership (MV), dateFounded, industry				
DBPedia	dbpedia.org/sparql	xml	location	43.720	5
	name, populationTotal, areaTotal, country, elevation				

Table 2.1: Overview of the datasets

(1) http://www.londonstockexchange.com/statistics/companies-and-issuers/companies-and-issuers.htm

2.2 LONDON STOCK EXCHANGE

The London Stock Exchange (LSE) offers a .csv file with a complete list of all listed companies (see (1)). This dataset consists of 2.396 entities containing mostly attributes describing financial information. While almost all attributes are complete, there are is one attribute (`Mkt Cap`), which gives the value 0 for some companies, indicating a lack of data or shares being traded for that company. In addition, this is a dataset containing official information by the LSE, compared to DBPedia and Freebase which contains information less complete and less trustworthy.

It may be possible to integrate the attributes `Sector` and `Sub sector` with the `industry` attributes of DBPedia and Freebase entities, depending on how much they overlap.

2.3 DBPEDIA

To access information about companies from DBPedia we plan to use the public SPARQL endpoint (at http://dbpedia.org/sparql). This also makes it possible to filter companies by certain attributes. An example would be to only use companies for this project that have the `numberOfEmployees` attribute with a value higher than 100. However, in this case this also reduces the number of entities available from DBPedia from 64.255 to 11.966. Using similar parameters for instances of locations, e.g. a `populationTotal` of at least 10.000, we reduce the number of entities from 725.546 to 43.783.

DBPedia also offers its data in the form of a data dump. However, here we may run into technical difficulties based on the size of the files (2.4GB compressed). Also, we would have to set up our own local server with a SPARQL endpoint.

2.4 FREEBASE

Freebase, like DBPedia, offers a web service which can be queried for data. Instead of SPARQL, however, the service is used by sending JSON requests. Like DBPedia this lets us select certain companies, or companies with certain attributes. By setting the same restrictions, e.g. only selecting companies with the `number_of_employees` attribute, the number of entities is reduced from 283.906 to 3.142.

An alternative would be to use a data dump they offer (developers.google.com/freebase/data), however here we run into similar size problems (22GB gzipped, 250GB uncompressed).

3 SATISFYING THE REQUIRED NUMBER OF ENTITIES

The first dataset, offered by the London Stock Exchange, contains listed companies on 28 August 2015. Based on on the number of companies (2.396) and the large number of locations accessible through DBPedia the integrated dataset should meet the basic requirement of having at least 2.500 entities.

Particularily the query based datasets (DBPedia and Freebase) are by nature flexible and allow us to select anywhere between several hundred and hundreds of thousands entities. As such an alternative would be to not restrict the target integrated dataset to companies listed in the LSE, and to use the LSE dataset for additional information on some companies.

Another option would be to choose a different stock exchange instead of the London Stock Exchange, i.e. one with more listed companies. An example would be NASDAQ, which offers a similar .csv file with fewer attributes per company but a larger number of entities (about 5.700).

Class Name	**Attributes Name**	**Datasets in which attribute is found**
company	(company)name	dataset 1,2,3
location/company	country	dataset 1,2,3,4
company	industry	dataset 2,3
company	revenue	dataset 2,3
company	number Of Employees	dataset 2,3
company	date founded	dataset 2,3
location/company	headquarters	dataset 2,3
company	Sector	dataset 1
company	Sub Sector	dataset 1
company	County of Incorporation	dataset 1
company	Market	dataset 1
company	Mkt Capacity	dataset 1
company	International Main Market	dataset 1
company	FCA Listing Category	dataset 1
company	list date	dataset 1
company	operating Income	dataset 2
company	key People	dataset 2,3
company	net Income	dataset 2
location	population total	dataset 4
location	area total	dataset 4
location	elevation	dataset 4

Table 4.1: Integrated target schema