Experiments with Wikipedia Cross-Language Data Fusion

Eugenio Tacchini[1], Andreas Schultz[2] and Christian Bizer[2]

[1] Università degli Studi di Milano,
Dipartimento di Tecnologie dell'Informazione,
Via Bramante 65-26013 Crema (CR), Italy
eugenio.tacchini@unimi.it

[2] Freie Universität Berlin
Web-based Systems Group
Garystr. 21 - D-14195 Berlin, Germany
aschultz@mi.fu-berlin.de, chris@bizer.de

Abstract. There are currently Wikipedia editions in 264 different languages. Each of these editions contains infoboxes that provide structured data about the topic of the article in which an infobox is contained. The content of infoboxes about the same topic in different Wikipedia editions varies in completeness, coverage and quality. This paper examines the hypothesis that by extracting infobox data from multiple Wikipedia editions and by fusing the extracted data among editions it should be possible to complement data from one edition with previously missing values from other editions and to increase the overall quality of the extracted dataset by choosing property values that are most likely correct in case of inconsistencies among editions. We will present a software framework for fusing RDF datasets based on different conflict resolution strategies. We will apply the framework to fuse infobox data that has been extracted from the English, German, Italian and French editions of Wikipedia and will discuss the accuracy of the conflict resolution strategies that were used in this experiment.

Keywords: Wikipedia, DBpedia, Web of data, data fusion, information quality evaluation

1 Introduction

Different Wikipedia language versions can describe the same type of objects in different ways, using different infobox templates[1], providing different and often conflicting information about the same topic. As an example, the English version of Wikipedia currently states that the city of Munich has a population of 1,356,594

1 http://en.wikipedia.org/wiki/Category:Infobox_templates

people while according to the German version the population of the same city is 1,315,476.

Handling these differences by applying data fusion algorithms can increase the information quality [1] of the resulting knowledge base if compared to the knowledge base derived from single Wikipedia editions: in the previous example it would be desirable for a user who enquires about the population of Munich to get the value provided by the German edition, which is more up-to-date, instead of the value provided by the English one. To get these results we need good heuristics which help recognize the correct dataset (Wikipedia edition) to choose from.

This paper examines the hypothesis that by extracting infobox data from multiple Wikipedia editions and by fusing the extracted data among editions it should be possible to complement data from one edition with previously missing values from other editions and to increase the overall quality of the extracted dataset by choosing property values that are most likely correct in case of inconsistencies among editions.

The work is structured as follows: we review related work in section 2; we give an overview of the DBpedia information extraction architecture, which we used to extract infobox data from different Wikipedia editions, in section 3; section 4 describes the data fusion framework that was used to merge data between editions. Section 5 presents the results of our experiments with applying different conflict resolution strategies to merge data between Wikipedia editions and estimates the accuracy of the fused datasets by comparing them to external “trusted” data.

2 Related Work

Data fusion is the process of merging multiple records representing the same real-world object into a single, consistent, and clean representation [3]. Beside of identity resolution, data fusion involves choosing the values that are most likely correct out of conflicting values within different data sets by applying conflict resolution strategies.

Data fusion is mainly addressed in the database research field. An overview of the field is given by Bleiholder and Naumann in [3].

In order to develop conflict resolution strategies for the Wikipedia use case, we reviewed existing work on Wikipedia information quality assessment. In [5] two metrics are used as a simple measure for the reputation of an article: Rigor (total number of edits of an article) and Diversity (total number of unique editors); the authors also verify that these measures positively change if an article gets a press citation. In [6] a rich citation-based trust evaluation is implemented. In [7] a list of quality metrics such as number of registered user edits, article length, currency, number of unique editors are applied to compute the quality of an article; as an experiment, the computed quality is used trying to recognize the *featured* Wikipedia articles. In [8] a Bayesian network model is used to compute the trust of an article, based on who edited the article (unregistered user, registered user or administrators) and on the status of the article (normal, to be cleaned, featured).

3 Infobox Data Extraction

The DBpedia project[2] extracts structured information from Wikipedia and makes this information available on the Web of Data [2]. Over 2.6 million Wikipedia entities are currently described in RDF [9], published according to the Linked Data principles [10, 11] and queryable via SPARQL [12].

Besides free text, Wikipedia articles contain structured information in the form of links, geo-coordinates, categories, links between different language versions of an article and infobox-templates which contain facts in a table like fashion. The aim of the DBpedia extraction framework[3] is to parse this information and to transform it into a consistent structural form, namely RDF. Wikipedia data dumps offered by the Wikimedia Foundation serve as the main data source in the extraction process.

We have used the DBpedia information extraction framework to extract infobox data from English, German, Italian and French editions of Wikipedia. The extraction framework also solves several problems that would otherwise hinder the fusion of the different language versions.

At first a common ontology has to be established for all participating sources. For this purpose we used the already existing DBpedia ontology[4] and we applied the mapping based approach of the extraction framework to the Wikipedia versions which we needed for our experiments. The main idea is to map infobox-templates coming from different Wikipedia editions which describe the same concept to the same class of the ontology and to map template properties to ontology properties.

The next step is to establish unique URIs of resources among the different Wikipedia editions; this step can be seen as the linking or duplicate detection step done in data integration. The employed approach to generate DBpedia URIs is to take the unique article name and prepending the DBpedia specific namespace (http://dbpedia.org/resource/) to it; however, article names can differ among language editions, so the Wikipedia interlanguage links[5] are exploited to identify every resource by the URI of its English-edition equivalent (if existent) in order to achieve unique identifiers.

An advantage of the afore-mentioned mapping-based approach is that it handles a third problem regarding data fusion, the representation of literals. Literals can be found in all kinds of formats and units, strongly depending on the language edition. The canonization of these values is part of the DBpedia extraction framework and facilitates an easier handling in the fusion process.

For our experiments the following language versions of Wikipedia were extracted and mapped to the common ontology: German, Italian and French, while the English version was already mapped to the DBpedia ontology and extracted[6].

---

2 http://wiki.dbpedia.org/

3 http://wiki.dbpedia.org/Documentation

4 http://wiki.dbpedia.org/Ontology?v=1cwu

5 http://en.wikipedia.org/wiki/Help:Interlanguage_links

6 http://download.wikimedia.org/backup-index.html Versions of the Wikipedia dumps: en 08.10.2008, de 11.10.2008, it 30.10.2008, fr 17.10.2008

4 Data Fusion Framework

We have developed a framework for fusing RDF data from local and remote RDF data sources. The framework conducts two basic steps: 1. query each source to get the required data, and 2. apply a strategy to merge the data from the different sources. Strategies apply different heuristics and thus lead to different outcomes. Provenance information in the resulting dataset is preserved by distributing the outcome to different named graphs [13], one for each source.

For our experiments we developed several strategies for the complementation and conflict resolution of the source data, ranging from a simple union with duplicate elimination to quality based conflict resolution. Table 1 and 2 summarize the strategies. The strategies are partitioned in augmentation and quality related strategies. The goal of the former is solely a quantitative one, whereas the latter, choosing on the base of a quality evaluation process, focuses on increasing the quality of the resulting dataset, albeit they often also augment it. It should be noted that in every single execution of a strategy the data for one property of exactly one entity of the specified class is processed.

Table 1. Augmentation based strategies

Onevalue	This strategy chooses the first value it comes across only. A check order of the sources can be defined.
Union	All values from all sources are taken for the resulting dataset.

Table 2. Quality based strategies

Democratic	The choice is based on the number of sources which share the same value for the same object instance/property couple. It is also possible to assign a weight for each source.
Geographic	The choice is based on the provenance information of the examined entity.
Edits number	The choice is based on the number of edits a page has received since its creation.
Filtered edits number	Same as above but the edits marked as "Minor" by the users are not taken into consideration.
Unique editors number	The choice is based on the number of different users who edited a page since its creation.
Accesses number	The choice is based on the number of visits a page has received since its creation or in general since a starting date
Last update date time	The choice is based on the date and time of the most recent edit a page has received

We will explain the quality based strategies in more detail as follows:

Democratic: This strategy is useful if many sources exist and/or the data of these sources overlap to a high degree. All candidate values are handled like in a majority

decision: the value that gets the most votes - in this case, appears in the most sources – will be chosen. Additionally the user can define a weight for each source, that affects the ranking.

Geographic provenance strategy: The idea behind this strategy is the assumption that the information of concepts that are localized - like cities for example - is better maintained by people who are located near this concept. The term “location” could also be expanded in a broader or more abstract sense like “intellectual proximity” or “cultural proximity”. For this strategy it has to be clearly defined how to categorize the entities and the sources, so the information is chosen by the source that falls into the same category of the entity. An example is to categorize cities by their "country property" (e.g. locatedIn) and choose the information from the source of the same country, in our case the suitable DBpedia language edition.

Wikipedia based quality strategies: The Wikimedia Foundation and other institutions offer metadata[7] for each page that include various statistics gathered about the changes that occur. The idea is to use these statistics to compute a quality ranking of the data from different sources, in our case, for the different language versions of an article in DBpedia. So this is a Wikipedia/DBpedia specific strategy. Table 2 shows all the implemented approaches for computing scores from this metadata which could be alternatively chosen; for the first four cases holds: the higher the number, the higher the score; for the last one, pages having more recent updates get higher score.

The developed data fusion famework provides for applying different fusion strategies to different properties of an entity. All aspects of the fusion process can be defined in a XML configuration file. The different configuration options are explained in the following.

The data sources are defined under the element *source* as shown in the example below:

```
<source id="dbpedia-en" type="sparql-endpoint"
augment="true">

  <url>http://localhost/sparql</url>

  <graph>dbpedia-en</graph>

</source>
```

7 sources of the Wikipedia quality indicators:
- Accesses numbers: http://wikistics.falsikon.de/dumps.htm (July, August and September)
- Other indicators:
http://download.wikimedia.org/itwiki/20081030/itwiki-20081030-stub-meta-history.xml.gz
http://download.wikimedia.org/enwiki/20081008/enwiki-20081008-stub-meta-history.xml.gz
http://download.wikimedia.org/dewiki/20081206/dewiki-20081206-stub-meta-history.xml.gz
http://download.wikimedia.org/frwiki/20081201/frwiki-20081201-stub-meta-history.xml.gz

The *id* attribute is the unique name of the source; the *type* characterizes the access method that, in this version, is limited to SPARQL-endpoints. The optional augment *attribute*, if set for one source, makes sure that entities from other sources not present in the augmented one will be ignored. This attribute was set for the English DBpedia dataset for all our experiments: an entity from a non-English dataset not available in the English dataset was therefore ignored.

Besides attributes source-elements have two sub-elements: *url* and *graph*, which define the URL of the SPARQL-endpoint and optionally the named graph containing the data.

An optional default setup of fusion strategies is possible under the element *strategy-config* and can be used to set default configurations for each strategy for later reuse. Such a definition of a strategy element has the following structure:

```
<strategy-config>

 <strategy type="single-value" name="democratic">

    <args>

           <arg id="dbpedia-en" value="3" />

           <arg id="dbpedia-de" value="2" />

           <arg id="dbpedia-it" value="2" />

           <arg id="dbpedia-fr" value="2" />

    </args>

 </strategy>

 …

</strategy-config>
```

Types can be set to *single-value* or *set-value*, which practically means that for a specific property only one value per entity should be chosen or a set of values. Birth date of a person is an example for the single value case, whereas band members would be a candidate for the set-value case. An optional *args* element defines the strategy arguments in an associative array fashion and is used to set up the strategy. Fusion strategies are applied to properties of specific classes:

```
<class URI=”http://dbpedia.org/ontology/Film”>

  <property URI=" http://dbpedia.org/ontology /runtime">

    <strategy type="single-value" name="democratic" />

  </property>

</class>
```

In this case no arguments are supplied to the strategy and in this way the default configuration - only if defined beforehand - is used.

5 Experiments

In order to test our framework, we applied it to different classes of objects extracted from Wikipedia infoboxes, selecting specific properties for each class. We evaluated the information quality of our resulting dataset comparing it with the information extracted from sources external to Wikipedia which we assume to be accurate. As our goal was to improve the English dataset (which is the one currently used by DBpedia to answer queries), the same evaluation was also performed on this dataset only (without applying data fusion); in this way we could verify if the fusion process impacted positively on the information quality level. This is the general approach we used for the experiments, in order to easily get the results, for some classes additional or different steps were done. Three of the experiments we did are described in details in the following paragraphs.

5.1 Dataset augmentation using a simple UNION operator

The first experiment focused on the use of the union operator applied to object properties. We chose to extract the starring property of Wikipedia articles about movies. The strategy was thus to just merge starring information coming from different Wikipedia editions in order to produce a resulting movies-actors dataset which was more complete than the one provided by the English edition only.

We extracted, using the DBpedia extraction framework, the value of the starring property for all the articles that used the “Infobox Film” template in the English version. Analogue templates are used by the German (Infobox Film), Italian (Film) and French (Infobox Cinéma (film)) versions; all the infobox templates included the starring property and this allowed for extracting its value from the four different language versions.

At the end of the process we managed to extract starring information for 30,390 movies and 118,897 movie-actor triples for the English version of Wikipedia, 7,347 movies and 42,858 movie-actor triples from the German version, 6,759 movies and

31,022 movie-actor triples from the Italian version, 1,171 movies and 3,739 movie-actor triples from the French version.

We then used our framework to produce a new dataset of movies which includes the starring values from all four starting dataset and we get a dataset composed by 143,654 movie-actor triples, augmenting the English dataset by 20.82%.

We then created a dataset composed only of the movie-actor triples added to the English dataset and compared this dataset with the IMDB database[8], which provides, among other data, for each movie, the list of actors who played a role in it.

In order to link DBpedia extracted movies and actors with the corresponding IMDB entries we used movie titles and actor names. In this example the linking process couldn't be accurately done like in the following experiments because movie titles and actor names in the IMDB dataset are not unique and are expressed in a format that differs from the one of DBpedia.

After this linking procedure we got 11,224 movie-actor triples and 61% of them are positively verified by the IMDB database check. The result of the experiment is positive because we expanded the dataset and most of the movie-actor triples were correct.

5.2 Data fusion using different information quality indicators

The second experiment we did focused on the use of the Wikipedia-based information quality indicators implemented in the framework. We took into consideration Wikipedia articles about minor planets; in particular we extracted the values of the orbital eccentricity property. This is a property whose values we could check from the MPC Orbit (MPCORB) Database[9], a public database which contains orbital elements for more than 200,000 minor planets. The strategy was thus to fuse information coming from different Wikipedia editions using some of the implemented quality indicators proposed in literature in order to produce a resulting planets dataset whose information quality was higher than the one provided by the English edition only, i.e. whose orbital eccentricity values were closer to the ones provided by the MPCORB database.

We extracted, using the DBpedia extraction framework, the value of the eccentricity property for all the articles belonging to the “planets” class i.e. the articles that uses the “Infobox Planet” template in the English version. Analogue, though not identical, templates are used by the German (Infobox Asteroid), Italian (Corpo celeste) and French (Infobox Planète mineure) versions; all the infobox templates included the eccentricity property and this allowed to extract its value for the four different language versions.

At the end of the process we managed to extract eccentricity information for 11,682 planets from the English version of Wikipedia, 11,251 planets from the Italian version, 2,502 planets from the German version and 267 planets from the French version.

8 ftp://ftp.fu-berlin.de/pub/misc/movies/database/, data retrieved 2009 Feb. 23

9 http://www.cfa.harvard.edu/iau/MPCORB.html, vers. 2009 Feb. 5

The subset of planets we took into consideration for the experiment was composed of all the planets included in both the English and MPCORB dataset and at least in one of the other datasets. In order to link DBpedia extracted planets with the MPCORB planets we used the name of the planet, which is unique. The final dataset was composed by 11,033 planets.

We then built an "ideal" selection of planets, choosing, for each planet, the data coming from the language version whose eccentricity value is closest to the one provided by the MPCORB dataset; this ideal selection was composed of 9.937 planets extracted from the English version, 962 from the Italian version, 127 from the German version and 7 from the French version. Using this selection it was possible to improve the quality of the data (measured as the sum of the absolute differences between the eccentricity value provided by DBpedia and the MPCORB's eccentricity value) by 17.47% in respect to a selection which just chose all values from the English version.

We then tried five different data quality indicators in order to see which one performed better and thus were able to create a selection which is as close as possible to the "ideal" selection; the results are shown in Table 3.

Table 3. Second experiment, performance of the information quality indicators tested

I.Q. indicator	Percentage of planets correctly selected
Edits number	10.16%
Filtered edits number	69.49%
Unique editors number	11.28%
Accesses number	19.43%
Last update date time	42.38%

The evaluation of the articles using the number of filtered edits is the one that performed better; 69.49% could be considered a good results but in this case, in which in more than 90% of the articles the information quality is higher in the English version (see "ideal" selection above), the final performance is worse in comparison to an approach which chooses all values from the English version so the final result for this experiment can't be considered positive. The information quality indicators proposed in most of the literature seem to work not very well, at least for this class of objects; one of the reason for poor performances of two of the edits-related indicators could be that we assumed that each edit operation added the same value to an article but, depending on author, size/type of content and other parameters the operation can increase (or, in some cases, decrease) the quality of an article at various levels. There are some parameters which can help us from this point of view (e.g. the minor parameter that we use for the filtered edits indicator) but we also have to take into consideration that the decision on marking an edit operation as minor is left to its author so in some cases the attribute could be unreliable.

5.3 Data fusion based on geographic provenance

The third experiment we did focused on the use of a promising information quality indicator: the geographic provenance. Our hypothesis was that for the class of objects that have a geographic provenance or localization (e.g. cities or people), data should be more accurate if taken from the Wikipedia version of the country they are related to. We took into consideration Wikipedia articles about cities; in particular we extracted the population data of Italian cities. We chose to focus on Italian cities because a public and up-to-date database providing data about cities population is available from the ISTAT (the national statistical institute of Italy) Web site[10].

We extracted, using the DBpedia extraction framework, the value of the population property for all the articles that used the "Infobox CityIT" template in the English version. The analogue template used for the Italian version (the only non-English version considered in this experiment) was "Comune".

In order to link DBpedia cities with ISTAT database's cities we used the ISTAT code, which is a unique identifier assigned to Italian cities; we got that code from the Geonames database dump[11] through the DBpedia - Geonames links dataset[12].

At the end of the process we managed to extract population information for 7,095 Italian cities from the English version of Wikipedia and 7,055 Italian cities from the Italian version.

The subset of cities we took into consideration for the experiment was composed of all the cities included in the English dataset and also in both the ISTAT and the Italian dataset. The final dataset was composed of 6,499 cities.

Following our initial hypothesis, we argued that Wikipedia articles about Italian cities were more accurate in the Italian Wikipedia Version. We thus compared population data of both the English and the Italian datasets with the data provided by ISTAT and these were the results: for 59% of the cities Italian data was more accurate (closer to ISTAT data); for 13% of the cities the quality was the same in both the datasets and for the remaining cities (28%) English data was more accurate. The final result for this experiment can be considered positive because the information quality of the resulting dataset is better in respect to an approach which chooses all values from the English version. This result confirmed our initial hypothesis; for articles with strong geographic localization characteristics as cities data should be more accurate if taken from the Wikipedia version of the provenance country.

10 http://demo.istat.it/bilmens2008gen/index02.html, data retrieved 2009 Feb. 13

11 http://download.geonames.org/export/dump/IT.zip, data retrieved 2009 Feb. 13

12 http://downloads.dbpedia.org/3.2/links/links_geonames_en.nt.bz2, data retrieved 2009 Feb. 16

6 Conclusions and Future Work

We presented the first version of a framework which is able to perform data fusion among different RDF datasets and which provides several conflict resolution strategies. We tested the framework in the Wikipedia/DBpedia domain, fusing data extracted from different Wikipedia language versions and we demonstrated that in some cases it is possible to increase the quality of the extracted information compared to extracting from the English Wikipedia edition only.
The results of the experiments were not always positive. As the quality of data within the English Wikipedia edition is already relatively high, it was difficult to improve data from the English edition with data from other editions. On the other hand, as the datasets that were extracted from other editions were relatively sparse, the solution proposed should work much better for augmenting a non-English Wikipedia version with the information extracted from the English version.

The geographic provenance of a DBpedia object is a promising indicator for quality evaluation, so one of the directions for future works will be an improvement of its implementation. An improvement of the other indicators is also desirable, especially in the direction of allowing to express the score with an higher level of granularity: as an example consider the possibility to have the last update date referred not to the whole page but to a fragment of it (a row of an infobox), this would give us the possibility to evaluate the currency of a single property instead of the currency of the page. We also have to proceed in the direction of fusing more Wikipedia editions; we tested our framework with four editions but adding other language versions can add more (potentially good) sources and in this way improve the information quality of the final dataset.

References

1. Bizer, C.: Quality-Driven Information Filtering in the Context of Web-Based Information Systems. PhD thesis, Freie Universität Berlin (2007)
2. Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R., Ives, Z.. Dbpedia: A nucleus for a web of open data. Proceedings of ISWC07 (2007)
3. Bleiholder, J. and Naumann, F. 2008. Data fusion. ACM Comput. Surv. 41, 1 (Dec. 2008), 1-41. DOI= http://doi.acm.org/10.1145/1456650.1456651
4. Naumann, F., Bilke, A., Bleiholder, J, Weis, M.: Data Fusion in Three Steps: Resolving Schema, Tuple, and Value Inconsistencies. IEEE Data Engineering Bulletin 29(2):21-31 (2006)
5. Lih, A.: Wikipedia as Participatory Journalism: Reliable Sources? Metrics for Evaluating Collaborative Media as a News Source. Proceedings of the Fifth International Symposium on Online Journalism (2004)
6. McGuinness, D. L., Zeng, H., Pinheiro da Silva, P., Ding, L., Narayanan, D., Bhaowal, M.: Investigations into trust for collaborative information repositories. Workshop on the Models of Trust for the Web (MTW'06) (2006)

7. Stvilia, B., Twidale, M. B., Smith, L. C., Gasser, L.: Assessing information quality of a community-based encyclopedia. In: Proceedings of the International Conference on Information Quality - ICIQ 2005. Cambridge, MA. 442-454 (2005)
8. Zeng, H., Alhoussaini, M.A., Ding, L., Fikes, R., McGuinness, D.L.: Computing trust from revision history. Intl. Conf. On Privacy, Security and Trust (2006)
9. Beckett, D.: RDF/XML Syntax Specification (Revised). W3C Recommendation. http://www.w3.org/TR/rdf-syntax-grammar/ (2004)
10. Berners-Lee, T.: Linked data. http://www.w3.org/DesignIssues/LinkedData.html (2006)
11. Bizer, C., Cyganiak, R, Heath, T.: How to publish linked data on the web, http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial/ (2007)
12. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF. W3C Recommendation. http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/ (2008)
13. Carroll, J., Bizer, C., Hayes, P., Stickler, P.: Named Graphs. Journal of Web Semantics, Vol. 3, Issue 4, p. 247-267 (2005)