Constructing a Comprehensive Knowledge Base from DBpedia

Ying Xu, Sep. 3rd, 2012
Department of Software Engineering
Southeast University & Monash University, Suzhou, China

***Abstract* - This report gives an overview of knowledge base construction based on Wikipedia and DBpedia. A brief solution is given by proposing a framework that can combine information from Wikipedia updates and Baidu Baike web pages into a localized DBpedia and problems of current situation are presented. Several disciplines including Web Crawler, Machine Learning, and Natural Language Processing are employed to enable the whole framework. Finally, a conclusion is provided to specify achievements and further improvements of this research.**

***Keywords* - Knowledge base, Semantic web, DBpedia, Wikipedia, Baidu Baike**

I. INTRODUCTION

Wikipedia has been mined by many previous researchers to extract knowledge because it is relatively more reliable and comprehensive compared to other web pages. In order to construct a machine-readable knowledge base, tedious work should be done to transform the text-based Wiki pages into structured format. Fortunately, with the development of semantic web (e.g. "Web of data"), knowledge can now be expressed in more structured way and can be easily accessed by machine. The knowledge base DBpedia [1] is one of the best choices for researchers. It is extracted from Wikipedia and has the inherent advantage derived from semantic web. It not only presents Wikipedia knowledge in a machine accessible way, but also connects entities with those of multiple external data sets to enable more complex queries to the web of data. The 35 data sets that are connected by DBpedia including OpenCyc, BBC Wildlife Finder, YAGO2, New York Times, WordNet Classes and so on. In fact, DBpedia is acting as one of the most important hub inter-connecting the semantic web. It in itself is a quite good knowledge base that has already support numerous of applications, like DBpedia Spotlight [2], DBpedia Mobile [3] and DBpedia Lookup. Using DBpedia as core and World Wide Web as the fertilize soil for knowledge growing can build up a strong and robust knowledge base.

II. OBJECTIVES

However, DBpedia can hardly satisfy the requirement of all knowledge base based applications. In fact, a localized knowledge base will perform better for specific target. As DBpedia is periodically published, a framework which automatically combines Wikipedia updates into localized DBpedia need to be developed to enable an update-to-date knowledge base. Furthermore, DBpedia only contain relatively comprehensive knowledge in English world. When comes to some country specific knowledge, local encyclopedia will give more satisfactory answer. As for Chinese applications, encyclopedia like Baidu Baike [4] may provide knowledge that is not available in DBpedia. Combining information from Baidu Baike to DBpedia will not only make up the knowledge gap, but also enrich multilingual edition of DBpedia. This paper proposed a framework setting up a localized update-to-date DBpedia based knowledge base, which extracts information from other data sets.

III. METHODOLOGY

DBpedia consists of three parts, ontology, data sets and links to other data sets. All the data is available in RDF triples, N quads and turtle format, so it can be easily stored in the Virtuoso RDF store. With the help of Virtuoso, our applications can visit DBpedia through SPARQL and reasoning can be performed through the inference engine.

3.1 Investigation

Wikipedia and Baidu Baike are the two encyclopedias that we plan to extract knowledge from. Wikipedia is the original source of DBpedia and Bizer [5] already provide a framework to automatically extract inforbox information from Wikipedia pages. However, the body of Wikipedia contains more valuable text information and page links which can be used to enrich the knowledge base. Lange [6] populates the Wikipedia inforbox by extracting attribute values from the article's text. Here we just use the results to populate DBpedia entities.

Baidu Baike was much more difficult because it has no well-formatted templates as that defined in Wikipedia. Fortunately, most web pages in Baidu Baike contain a table structure that demonstrates the basic information of specific category entity. For example, Movie entities always contain table giving following attributes: Chinese title, English title, publish time, publisher, directors, actors, scriptwriters, length and release date. Information contained in this table is structured and more information relies on the text description edit in different chapters.

3.2 Framework

The framework constructing a DBpedia based knowledge base is demonstrated in the Figure 1.

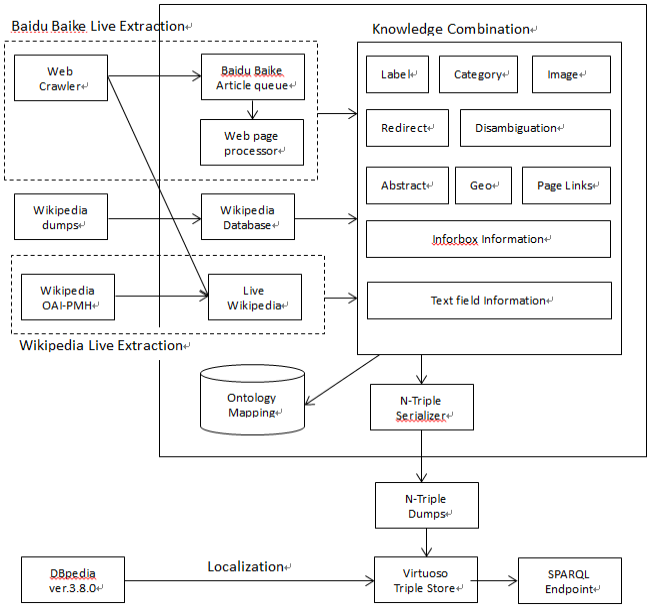

Fig. 1 Knowledge base construction framework

As mentioned before, DBpedia should be localized to support the subsequent enrichment process. Here we use the Virtuoso to store the ontology and datasets. Once the localization is successfully performed, the enrichment process can run SPARQL queries against its interesting part and updates could be made to the backend RDF dumps.

Wikipedia dump extraction. Monthly updates of the Wikipedia dumps are available from the Wikimedia Foundation. The dumps are SQL based and the extractor can easily acquire information using SQL queries.

Wikipedia live extraction. *Wikipedia OAI-PMH live feed* reports the change that has been made to Wikipedia instantly. The extractor extracts new RDF triples and update old ones to keep the knowledge base being coordinate with the latest Wikipedia updates.

Web crawler. Wikipedia and Baidu Baike pages can be visited by both web users and web crawlers. Because large amount of pages are contained in these two encyclopedias, a distributed web crawler is developed to avoid high access frequency to them, which will incur heavy traffic for the visited one.

Web page processor. The processor takes crawled web pages as input and computes validated RDF triples as output. Table elements are transformed directly into triples by defining the attribute name as predicate and the attribute value as object. Plan text are tokenized, lexically analyzed, grammatically analyzed and formulated as a Vector Space Model. Information required is extracted using the idea that specific information is more likely appear in the same contexts. Training set of text can be obtained from DBpedia itself or annotation.

Knowledge combination. The new RDF triples extracted from web pages should be consistent with the ontology defined in the DBpedia ontology specification. Overlapped ones are removed and updates are performed after consistency validation.

IV. Novelty

Till to now, various researches and applications are developed against it. However little of them try to make up its lack of non-English knowledge.

There's no doubt that Italian knows more about Italian villages. DBpedia workforce has realized this problem, and it tries to solve it by confusing knowledge from cross-language Wikis. On this way, quality of knowledge base is increased compared to that only depend on English Wikipedia edition. However, we believe that his method is limited by its source of knowledge. In fact, encyclopedias like Baidu Baike are more popular with Chinese people and much more accurate and comprehensive knowledge can be extracted from this kind of sources. Although DBpedia does excellent at linking entity to other semantic web datasets, none of the previous research has tried to explore to other encyclopedia web pages due to the difficulty of combining unstructured data with structured DBpedia. This is what we are trying to figure out in this research.

V. Conclusion

The knowledge base constructed shows that the web of data can connect data sources across the Internet and knowledge from local websites does better in provide localized information. Although DBpedia covers large scale knowledge in the world and can be updated by thousands of editors from all over the world, it can never know more about the Chinese villages than Chinese encyclopedias, such as Baidu Baike and Hudong. In this research, we successfully deploy DBpedia on local server and extract daily updates from Wikipedia to make it an up-to-date knowledge base. A framework is proposed to combine knowledge from other knowledge source in a consistent way. The entities and relations extracted from Baidu Baike better make up DBpedia's disadvantage in providing rich information about Chinese specific knowledge. The semantic search engine we developed on top of the knowledge base performs better results in term of precision ratio and recall ratio than just on top of DBpedia.

Further work will involve more data sources and more languages. How to enrich the DBpedia is in fact a policy issue. This research just focuses on digging specific knowledge buried in local encyclopedias in different countries. More problems will emerge if we use DBpedia in other aspects. For example, domain specific knowledge is better served by those professional websites such as Flickr.com, DBLP and BBC News; application oriented knowledge may come from calculation, reasoning or even human experts. No matter what DBpedia will be used to do, common problem of entity disambiguation, ontology matching and data consistency will always stand at the heart of knowledge base construction.

References

[1] *DBpedia*. Available from: http://dbpedia.org/.

[2] Mendes, P.N., et al., *DBpedia spotlight: shedding light on the web of documents*, in *Proceedings of the 7th International Conference on Semantic Systems*2011, ACM: Graz, Austria. p. 1-8.

[3] Becker, C. and C. Bizer, *DBpedia mobile: A location-enabled linked data browser*. Linked Data on the Web (LDOW2008), 2008.

[4] *Baidu Baike*. Available from: http://baike.baidu.com/.

[5] Bizer, C., et al., *DBpedia-A crystallization point for the Web of Data*. Web Semantics: Science, Services and Agents on the World Wide Web, 2009. **7**(3): p. 154-165.

[6] Lange, D., et al., *Extracting structured information from Wikipedia articles to populate infoboxes*. Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, ACM: Toronto, ON, Canada. p. 1661-1664.