Constructing Knowledge Base from DBpedia

Ying Xu

Southeast University-Monash University Joint Graduate School, Suzhou, China

Abstract. In this paper, we introduce a new direction of constructing up-to-date knowledge base. At first, we demonstrate the history and development of knowledge base. Then we review the conventional methods of constructing knowledge base from Wikipedia and several applications associated with the Semantic Web edition of Wikipedia: DBpedia, which provides machine-readable format of large scale knowledge in Wikipedia pages and can be combined with multiple datasets available on the Internet to form the so-called Web of Data. At last, conventional methods of extracting information from web pages are reviewed, which are used to complement DBpedia's disadvantage of time gaps between released editions.

Keywords: Knowledge base, Semantic web, Wikipedia, DBpedia, Information retrieval.

1. Introduction

Artificial Intelligence (AI), due to its inherited nature, can hardly be defined in a standard way. After more than 50 years of development from its foundation in 1956, several sub-problems are gradually derived and some of them are still very hot topics in the scholar field, for example, knowledge presentation, planning, learning, nature language processing, social intelligence and so on. Here we concentrate on knowledge presentation, which is the most basic problem of AI.

Knowledge Base (KB) is commonly considered as the place to store the knowledge extracted and to be used by the AI. It provides many AI applications with the extensive knowledge about the world. Most obvious example in our daily life is various search engines, which provide us with the links to the web pages containing things we want. In fact, behind each search engines we are familiar with, there exists a power knowledge base to satisfy increasing intelligent demands from web users. How to create a comprehensive and accurate knowledge base has become a hot research topic in the AI field.

The data source of KB can be an expert of specific field, electronic documents or the most popular World Wide Web. With the development of Internet, huge numbers of web pages containing knowledge are available on the web and projects like Wikipedia has provided high-quality knowledge to web users and researchers. In fact, if no conditions placed, Wikipedia itself is a comprehensive and up-to-date knowledge base. But unfortunately, this kind of knowledge base can only be read by human beings. It's not machine-readable. Computers feel more comfortable with structured data and that's why many KB developers spend years of time to convert unstructured web pages, including Wikipedia pages, to structured data.

Thanks to the emergence of semantic web, structured knowledge is available now on the web, among which DBpedia is the most famous. It is extracted directly from Wikipedia

and is described by Tim Berners-Lee as the one of the most famous part of the Linked Data projects. The things described by DBpedia include 764,000 persons, 573,000 places, 333,000 creative works, 192,000 organizations, 202,000 species and 5,500 disease [1]. Each part of the knowledge base can be used to support applications in a bunch of domains, e.g. health care, geography, entertainment, government affairs. In addition, DBpedia deals perfect with the relationships among the things it described, because all the relationships are standardized by the ontology defined by DBpedia. In another word, it provides with users a consistent dataset.

Despite of all the advantages of DBpedia, there still exist some problems need to be addressed. First of all, the DBpedia is periodically published and lot of knowledge may be lost during the release gap. Secondly, most information of DBpedia are extracted from the "infobox" of Wikipedia pages, which ignores rich information contained in the main body of description. Thirdly, outside the Wikipedia, there are many other encyclopedia web sites providing useful information, e.g. YAGO, Yahoo!, Baidu.

2. Scope and Method

Knowledge base construction is a complex task if accuracy, completeness and freshness are taken into consideration. It request collaborative knowledge from multiple subjects, such as Artificial Intelligence, Machine Learning, Data Mining, Semantic Web, Probabilistic Analysis and Nature Language Processing.

There are many sources of knowledge: DBpedia, Wikipedia, Other popular datasets (e.g. World Gazetteer) and web pages. First, detailed investigation should be done for all these knowledge sources. Then try to figure out how to combine them in a consistent way. Finally, applications like semantic search engine can be developed to evaluate the quality of the knowledge base.

3. Related Works

3.1 Wikipedia and DBpedia

Now we will review the development of Knowledge Base and its combining with Semantic web.

Knowledge representation and knowledge engineering are central to AI and many efforts have been made before to construct knowledge bases which can satisfy the demand of artificial applications. But it was until the Cyc project [2] when an ambitious and visionary target of knowledge base was formed. This project tried to codify millions of pieces of knowledge which contain not only the broad ontology scope but also detailed encyclopedia information of human common sense. This project was estimated by Doug Lenat saying that it will consume 350 men-years of effort [3].

But this estimation didn't freak out the knowledge base researchers. After the famous Tim Berners-Lee proposed a distributed hypertext system based on Internet protocols [4], Wikipedia, a free, collaborated edited and multilingual Internet encyclopedia was start to develop and gained attention from researchers. [5] was a hybrid knowledge base of structured and unstructured information extracted from Wikipedia, RDF data from DBpedia and other Linked Data resources. It support several applications like document concept prediction, cross document co-reference resolution, Entity Linking to KB entities and interpreting tables, which, in return contribute to enriching the knowledge base automatically. Wu [6] probes deeper into the ontology refinement with the help of Wikipedia Infobox data. This structured part of Wikipedia also used by [7] to enable users to ask complex questions

against Wikipedia knowledge. It provides another format of knowledge apart from structured database and RDF triples. In fact, knowledge can be represented in different format, including relational tables, RDF triples and as well as search interfaces. With the development of Wikipedia, many other encyclopedias are available on the Internet and many of them have some relationship with Wikipedia. Suchanek [8] presents a large ontology with high coverage and precision. It combines the category system of Wikipedia and the taxonomic relations from WordNet to form a consistent ontology. Although the English edition of Wikipedia is more comprehensive and informative than other language, KBs extracted from multilingual Wikipedia are also available over the net. But usually these KBs tend to combine the knowledge extracted from Wikipedia with those from local encyclopedias. Wang [9] tries to build a large-scale Chinese structured knowledge base from Chinese wiki resources, including Hudong and Baidu Baike.

It's obvious that Wikipedia has been regarded as a seed for growing knowledge and it is, but it has the disadvantage of full-text format, which limits the access from different users. Semantic web, first coined by Tim Berners-Lee [10], bring another spring to Knowledge engineering. The RDF triple imported from semantic web provide a new format to store great amount of knowledge. In addition, the standard ontology and logic constraints defined in RDFS make it possible to maintain consistency on large scale knowledge. DBpedia, a project aiming to extract structured content from the information created as part of the Wikipedia project, was available on the Internet from 2007 [11]. Bizer [11] and Auer [12] give the overall introduction to the project DBpedia, presenting the framework of transforming Wikipedia into RDF triple, several Semantic Web applications which consume DBpedia a main datasets and the linking relation from DBpedia to other datasets. From then on, a new field of vision was opened and many researchers begin to pay attention to this promising project. Torres [13] states an interesting work which adds complex semantic query results into Wikipedia. This paper propose a path indexing algorithm (PIA) find the best representative path in Wikipedia according to the result sets derived from semantic queries. Szczuka [14] uses DBpedia as a standard concept system which can be matched to the vector representation of test corpus. There are many other application situations where DBpedia performs well as a knowledge base. But there are several issues which should be addressed for us to better utilize this valuable knowledge base. Virtuoso provides a method to localize the published DBpedia datasets. It is a middleware and database engine hybrid which supports RDF storage, logical reasoning and SPARQL query access which deals with the storage issue of DBpedia. Kontokostas [15] takes the Greek DBpedia as an example to demonstrate how to generate, maintain and properly interlink language-specific DBpedia editions, which support multilingual access. At last, Hellmann [16] tries to complement DBpedia's disadvantage as a heavy-weight extraction process. It raises a live extraction framework to capture Wikipedia's tens of thousands of changes per day to make the releasing an up-to-date edition.

3.2 Traditional technologies

Unfortunately, few works focus on using DBpedia to construct a more comprehensive and accurate knowledge base. In fact, to address this problem, we have to go back to the original knowledge extraction from plain web pages. But this time, with DBpedia as the

core of the final KB, much tedious work will be skipped over and in return. Traditional technologies used to extract knowledge from web pages include Web Crawler, Nature Language Processing, and Name Entity Recognition. In addition, because DBpedia is stored as consistent Semantic Web datasets, we will employ ontology matching and several reasoning engines to satisfy the consistency demand of Semantic Web facts.

Web Crawler is an agent that browses the World Wide Web and downloads interested web pages according to the policy defined by the author. Here, it periodically extracts web pages which are used to update the knowledge base. Shkapenyuk [17] described and design a distributed web crawler which aims at interacts with millions of hosts over a period of time.

On the other aspect, the DBpedia already gives the ontology definition in different areas, e.g. person, organization, places, species and so on. Knowledge from web pages should be mapped to concepts and relations defined in the interested ontology. Craven [18] developed a trainable information extraction system, which takes the already exist ontology system and training data from hypertext as inputs and learns to extract knowledge from other web pages and hyperlinks on the Internet. Several learning algorithms are used to train the knowledge extraction system. One of the most significant problems of information extraction is Name Entity Recognition. Cucerzan [19] combined four classifiers (e.g. robust linear classifier, maximum entropy, transformation-based learning, and hidden Markov model) to recognize name, location and person with a performance of 91.6F on the English development data. A more detailed problem of entity disambiguation was addressed in [20]. This paper disambiguates entities through two factors. The first is the agreement between the contextual information extracted in Wikipedia and the contextual information in the training document. Another is the agreement among the category tags associated with the candidate entities. Here, getting accurate tags associated with entities was another problem need to be solved. Finally, we need to combine the ontology from DBpedia and the ones extracted from web pages. Noy [21] proposed an algorithm to semi-automatically merge ontology from different sources and determines the possible inconsistency in the state of ontology.

4. Conclusions

We can conclude that with the DBpedia as the core of the knowledge base, much tedious work will be skipped. We don't have to construct the knowledge base from the start. And the well-defined ontology given by DBpedia will standardize the process of extraction knowledge from the World Wide Web. In fact, numerous of other data sources apart from DBpedia are also available on the Internet. If we combine these data sources first and then use the Wikipedia and web pages as an up-to-date source of new knowledge, the quality of the knowledge will increase a lot.

As DBpedia is relatively new for many knowledge base engineers, little work focuses on constructing a knowledge base from the DBpedia. Filling the time gap between released editions is closely related to the conventional method of information retrieval from the web pages. Future work will emphasize on how combine the knowledge from DBpedia and the knowledge from web pages.

Another important issue that needs to be discussed is the multilingual edition of the knowledge base. Much of the works mentioned above are just available for English edition. How to enable interoperability among multiple languages and finally allow multilingual access from different users is a problem to be addressed in the future.

References

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

[2] Elkan, C. and R. Greiner, *Building large knowledge-based systems: representation and inference in the Cyc project.* Artificial Intelligence, 2006. 61(1): pp. 41-52.

[3] The Editors of Time-Life Books (1986). Understanding Computers: Artificial Intelligence. Amsterdam: Time-Life Books. p. 84. ISBN 0-7054-0915-5.

[4] Berners-Lee, T., *Information management: A proposal.* 1989.

[5] Syed, Z.S., *Wikitology: A novel hybrid knowledge base derived from wikipedia*, 2010, University of Maryland, Baltimore, USA.

[6] Wu, F. and D.S. Weld. *Automatically refining the wikipedia infobox ontology*. 2008. ACM.

[7] Hahn, R., et al., *Faceted Wikipedia Search*, W. Abramowicz and R. Tolksdorf, Editors. 2010, Springer Berlin Heidelberg. pp. 1-11.

[8] Suchanek, F.M., G. Kasneci, and G. Weikum, *YAGO: A Large Ontology from Wikipedia and WordNet.* Web Semantics, 2008. **6**(3): pp. 203-217.

[9] Wang, Z., J. Li, and J.Z. Pan, *Knowledge extraction from Chinese wiki encyclopedias.* Journal of Zhejiang University-Science C, 2012. 13(4): pp. 268-280.

[10] Lee, T.B., J. Hendler, and O. Lassila, *The semantic web.* Scientific American, 2001. 284(5): pp. 34-43.

[11] 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): pp. 154-165.

[12] Auer, S., et al., *Dbpedia: A nucleus for a web of open data.* The Semantic Web, 2007: pp. 722-735.

[13] Torres, D., et al. *Improving Wikipedia with DBpedia*. 2012. ACM.

[14] Szczuka, M., A. Janusz, and K. Herba, *Semantic Clustering of Scientific Articles with Use of DBpedia Knowledge Base.* Intelligent Tools for Building a Scientific Information Platform, 2012: pp. 61-76.

[15] Kontokostas, D., et al., *Internationalization of Linked Data: The case of the Greek DBpedia edition.* Web Semantics: Science, Services and Agents on the World Wide Web, 2012.

[16] Hellmann, S., et al., *Dbpedia live extraction.* On the Move to Meaningful Internet Systems: OTM 2009, 2009: pp. 1209-1223.

[17] Shkapenyuk, V. and T. Suel. *Design and implementation of a high-performance distributed web crawler*. 2002. IEEE.

[18] Craven, M., et al., *Learning to construct knowledge bases from the World Wide Web.* Artificial Intelligence, 2000. 118(1): pp. 69-113.

[19] Florian, R., et al. *Named entity recognition through classifier combination.* 2003. Association for Computational Linguistics.

[20] Cucerzan, S. *Large-scale named entity disambiguation based on Wikipedia data*. 2007.

[21] Noy, N.F. and M.A. Musen. *Algorithm and tool for automated ontology merging and alignment.* 2000.