Semantic Data Extraction from Infobox Wikipedia Template

Amira Abd El-atey
Faculty of computers and Information
Menofia University

Sherif El-etriby
Faculty of computers and Information
Menofia University

Arabi S. kishk
Faculty of computers and Information
Menofia University

ABSTRACT

Wikis are established means for collaborative authoring, versioning and publishing of textual articles. The Wikipedia for example, succeeded in creating the by far largest encyclopedia just on the basis of a wiki. Wikis are created by wiki software and are often used to create collaborative works. One of the key challenges of computer science is answering rich queries. Several approaches have been proposed on how to extend wikis to allow the creation of structured and semantically enriched content. Semantic web allows of creation of such web. Also, Semantic web contents help us to answer rich queries. One of the new applications in semantic web is DBpedia. DBpedia project focus on creating semantically enriched structured information of Wikipedia. In this article, we describe and clarify the DBpedia project. We test the project to get structured data as triples from some Wikipedia resources. We clarify examples of car resource and Berlin resource. The output data is in RDF (Resource Description Framework) triple format which is the basic technology used for building the semantic web. We can answer rich queries by making use of semantic web structure.

General Terms

Information retrieval; semantic web.

Keywords

Wikipedia ; semantic web ; DBpedia; data extraction framework; structured knowledge; wikipedia templates; media wiki software; infobox template.

1. INTRODUCTION

The free encyclopedia Wikipedia has been tremendously successful due to the ease of collaboration of its users over the Internet. The Wikipedia wiki is the representative of a new way of publishing and currently contains millions of articles. Wikipedia is a free, web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation. Its 18 million articles (over 3.6 million in English) have been written collaboratively by volunteers around the world and almost all of its articles can be edited by anyone with access to the site. Wikipedia was launched in 2001 by Jimmy Wales and Larry Sanger and has become the largest and most popular general reference work on the Internet, having 365 million readers. It is a natural idea to exploit this source of knowledge. Wikipedia has the problem that its search capabilities are limited to full-text search, which only allows very limited access to this valuable knowledge base [13]. The DBpedia project focuses on the task of converting Wikipedia content into structured knowledge, such that Semantic Web techniques can be employed against it, asking sophisticated queries against Wikipedia and linking it to other datasets on the Web. The project was started by people at the Free University of Berlin and the University of Leipzig, in collaboration with OpenLink Software and the first publicly available dataset was published in 2007. It is made available under free licences, allowing others to reuse the dataset [14].

Until March 2010, the DBpedia project was using a PHP-based extraction framework to extract different kinds of structured information from Wikipedia. This framework has been superseded by the new Scala-based extraction framework and the old PHP framework is not maintained anymore. The superseded PHP-based DBpedia information extraction framework is written using PHP 5. The new frame work written using Scala 2.8 is available from the DBpedia Mercurial (GNU GPL License). Wikipedia articles consist mostly of free text, but also include structured information embedded in the articles, such as "infobox" tables, categorization information, images, geo-coordinates and links to external Web pages. This structured information is extracted and put in a form which can be queried [4,8].

Semantic web is able to describe things in a way that computers can understand. Answering semantically rich queries is one of the key challenges of semantic web today [7]. In this article, we give an overview of Wikipedia and semantic web in section 2 and 3 respectively. Section 4 gives discusses structured data extraction framework of Wikipedia. Section 5 also discussed integration of semantic web data on the web. Section 6 gives overview about related work. Section 7 concludes and outlines conclusion and future work.

2. WIKIPEDIA

Wikipedia articles consist mostly of free text, but also contain different types of structured information, such as infobox templates, categorization information, images, geo-coordinates and links to external Web pages. This structured information can be extracted from Wikipedia and can serve as a basis for enabling sophisticated queries against Wikipedia content. DBpedia project extracts this structured information from Wikipedia and turns it into a rich knowledge base. This knowledge base can be used later to ask sophisticated queries. In this section we give an overview of MediaWiki templates and infobox template.

2.1 MediaWiki Templates

MediaWiki supports templates for Wikipedia by using MediaWiki software. The MediaWiki software is an open source software that wikiHow, Wikipedia, Wiktionary, and many other wiki sites are based upon. The wiki engine enables each member to search, read, add and edit articles, and thus improve the content of the wiki. Wiki software can be downloaded as a ready-made tool and in the majority of cases its use is free of charge [10].

MediaWiki supports a sophisticated template mechanism to include predefined content or display content in a determined way. Some of these MediaWiki templates are input box, message box and infobox. Infobox template is intended as a meta-template.

2.2 Infobox template

A special type of templates is infobox, aiming at generating consistently-formatted boxes for certain content in articles describing instances of a specific type. An infobox template is a fixed-format table designed to be added to the top right-hand corner of articles to consistently present a summary of some unifying aspect. An example of infobox template code is shown in Figure 1. It is about AlMenoufiya. As we see, the infobox is enclosed with {{ }} operators. Summary about Al Menoufiya is described as label/data rows. It describes data about Al Menoufiya such as name, image, country, area, population and other data about Al Menoufiya city. Infobox templates are used on pages describing similar content [1, 3]. The generated view of infobox code is shown in Figure 2. As we see, it visualizes the code we have described in Fig1. Other examples include Geographic entities, education, plants, organizations, people and so on.

```
{{Infobox
|name               = Al Menoufiya
|image              = Menofia.png
|country            = Al Menoufiya
|area             = 2554
|population         = 1780153
|population as of     = 1996
|population density   = 1088
|administration area = shbein elkom
|postal code         = 23511 – 23754
|lat_deg            = 30
|lat_min             = 26
|lat_hem             = North
|lon_deg             = 31
|lon_min              = 4
|lon_hem              = East
|Website             = [www.monofeya.gov.eg]
}}
```

Figure 1. Infobox template code Al Menoufiya

Al Menoufiya

Figure 2. Visualization of Infobox Template about Al Menoufiya

Country	Egypt
Area	2554km^2
Population	2,780,153(1996)
Population density	1088/Km^2
Administration area	Shbein elkom
Postal code	23511-23754
Coordinates	30^{o} 26^{1} 19^{11} North, 31^{o} 04^{1} 08^{11} East
Website	www.menofiya.gov.eg

3. SEMANTIC WEB

Semantic web is a new vision of current web. It is a web that is able to describe things in a way that computers can understand. it is not about links between web pages. The Semantic Web describes the relationships between things (like A is a part of B and Y is a member of Z) and the properties of things (like size, weight, age, and price). It has more standard and unstandard technologies. In this section we describe important semantic web technologies which are RDF and SPARQL.

3.1 The Resource Description Framework technology

The RDF (Resource Description Framework) is a language for describing information and resources on the web. Putting information into RDF files, makes it possible for computer programs ("web spiders") to search, discover, pick up, collect, analyze and process information from the web. The Semantic Web uses RDF to describe web resources. RDF usually displayed as A Subject-Predicate-Object. If you used RDF for representing data, you need a way for accessing information that mirrors the flexibility of the RDF information model. RDF query languages such as SPARQL query language.

RDF model for the web can be considered as the equivalent of the ER (Entity-Relationship) model for the RDBMS (relational database management system). Let’s look at a simple example. Consider the fact that “The book was written by Jane Doe.” In a traditional ER model, this information would be expressed as shown in Figure 3.A. An RDF graph of the same concept would be represented as in Figure 3.B. The RDF graph represents a node for the subject “book”, a node for the object “author”, and an arc/line for the predicate relationship between them. The goal of both an ER model and an RDF graph is to provide a human-readable picture that can then be translated into machine-readable format. Where a physical model in the relational database world can create DDL (data definition language) to execute on a database, an RDF graph can be translated into “triples” where each node and predicate is represented by a URI (uniform resource identifier) which provides the location of the information on the web network. For example, the above RDF graph can be represented as a triple shown in Figure3.C.

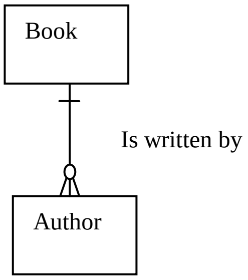

Figure 3.A. ERD model

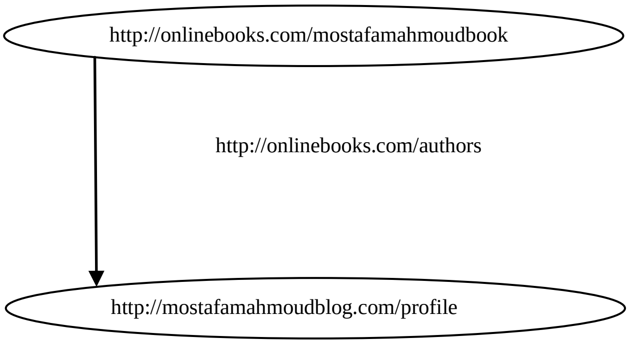
http://onlinebooks.com/mostafamahmoudbook
http://onlinebooks.com/authors
http://mostafamahmoudblog.com/profile

Figure 3.B. RDF model

```
<http://onlinebooks.com/mostafamahmoudbook>
<http://onlinebooks.com/authors><http://mostafa
mahmoudblog.com/profile>
```

Figure 3.C. RDF triple

3.2 SPARQL Protocol and RDF Query Language

SPARQL is query language for RDF documents. It is like querying database. We use SQL for querying database and SPARQL for querying RDF data. Semantic web provides a SPARQL endpoint for querying web data or knowledge base. Client applications can send queries over the SPARQL protocol to the endpoint. In addition to standard SPARQL, the endpoint supports several extensions of the query language that have proved useful for developing client applications, such as full text search over selected RDF predicates. We conduct simple query on simple RDF triple. This simple example shows a SPARQL query to find the author of a book from the given data graph. The query is shown in Figure 4. The query consists of two parts: the `SELECT` clause identifies the variables to appear in the query results, and the `WHERE` clause provides the basic graph pattern to match against the data graph. The basic graph pattern in this example consists of a single triple pattern with a single variable (`?author`) in the object position. This simple query means what is the author of mostafamahmoudbook. The query result will be http://mostafamahmoudblog.com/profile.

```
SELECT ?author
WHERE
{
<http://onlinebooks.com/mostafamahmoudbook>
```

Figure 4. RDF query of RDF triple.

4. STUCTURED DATA EXTRACTION FRAMEWORK

The DBpedia community uses a framework to extract different kinds of structured information from Wikipedia. The DBpedia extraction framework is written using Scala 2.8 under GNU GPL License. Some information in MediaWiki page is in structured form, which used infobox template and some cached in database, then there exist two methods to extract semantic relationships. First, map relationships that are already in database tables onto RDF [2] and second, we extract additional information directly from the article texts and infobox templates within the articles [1].

4.1 Extracting structured data from Wikipedia page

To reveal semantics in templates we follow some steps:

- Extract infobox template: infobox template on a Wikipedia page extracted by means of a recursive regular expression. Infobox template started with {{symbol and ends with the same symbol
- Parse template and generate appropriate triples: URL derived from the title of the Wikipedia page the template occurs in is used as subject. Each template attribute corresponds to the predicate of a triple and the corresponding attribute value is converted into its object.

Resources in Wikipedia are assigned a URI according to the pattern http://dbpedia.org/resource/Name, where Name is taken from the URL of the source Wikipedia article, which has the form http://en.wikipedia.org/wiki/Name.

In the following results, we define the resource also, display different extractors for that resource in RDF format. RDF usually displayed as A Subject-Predicate-Object (SPO), subject is a resource, predicate is a resource, and object is a literal [17].

We drive two examples to car resource and Berlin resource of Wikipedia. Data extracted are shown as triples, which are subject, predicate and an object. Output RDF descriptions of car resource shown in Table I. DBpedia resource identifier http://dbpedia.org/resource/Car set up to return RDF descriptions when accessed by Semantic Web agents. Car resource on English Wikipedia redirected to automobile resource, this is obviously shown in abstract extractor. Output RDF descriptions of Berlin resource are shown in Table II. DBpedia resource identifier, http://dbpedia.org/resource/ Berlin set up to return RDF descriptions when accessed by Semantic Web agents. Berlin resource on English Wikipedia has no redirection, so in abstract extractor the text displayed in Wikipedia similar to the text displayed in object literal.

In car resource, we clarify label extractor with its S (subject), P (predicate), and O (object). Also, we clarify wiki page extractor with its S (subject), P (predicate), and O (object). Also, we clarify long abstracts extractor with its S (subject), P (predicate), and O (object).

In Berlin resource, we clarify label extractor with its S (subject), P (predicate), and O (object). Also, we clarify wiki page extractor with its S (subject), P (predicate), and O (object).

TABLE I. Output RDF descriptions of car resource

Label extractor	
S	http://dbpedia.org/resource/car
P	http://www.w3.org/2000/01/rdf-schema#label
O	"car" (xml:lang="en")

Wiki page extractor	
S	http://dbpedia.org/resource/car
P	http://xmlns.com/foaf/0.1/page
O	http://en.wikipedia.org/wiki/car

Long abstracts extractor	
S	http://dbpedia.org/resource/car
P	http://dbpedia.org/property/abstract
O	"#REDIRECT Automobile" (xml:lang="en")

TABLE I. Output RDF descriptions of Berlin resource

Label extractor	
S	http://dbpedia.org/resource/Berlin
P	http://www.w3.org/2000/01/rdf-schema#label
O	"Berlin" (xml:lang="en")

Wiki page extractor	
S	http://dbpedia.org/resource/Berlin
P	http://xmlns.com/foaf/0.1/page
O	http://en.wikipedia.org/wiki/Berlin

5. INTEGRATION OF DBPEDIA DATA ON THE WEB

DBpedia is served on the Web under the terms of the GNU Free Documentation License. The DBpedia knowledge base can be access through some mechanisms.

5.1 SPARQL Endpoint

We provide a SPARQL endpoint for querying the DBpedia knowledge base. Client applications can send queries over the SPARQL protocol to the endpoint at http://dbpedia.org/sparql. In addition to standard SPARQL, the endpoint supports several extensions of the query language that have proved useful for developing client applications, such as full text search over selected RDF predicates, and aggregate functions. To protect the service from overload, limits on query complexity and result size are in place. The endpoint is hosted using Virtuoso Universal Server [6].SPARQL allows users to write globally unambiguous queries. For example, the following query returns names and emails of every person in the world:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?email
WHERE {
?person a foaf:Person.
  ?person foaf:name ?name.
  ?person foaf:mbox ?email.
}
```

5.2 Interlinking DBpedia with other Open Datasets

The DBpedia dataset is interlinked with other data sources on the Web using RDF links. Overview of some linked data sources shown in Figure 5. The DBpedia interlinking effort is part of the Linking Open Data community project of the W3C Semantic Web Education and Outreach (SWEO) interest group. RDF links enable query to navigate from data within one data source to related data within other sources using a Semantic Web browser [5, 16]. The DBpedia data set is interlinked with various other data sources.

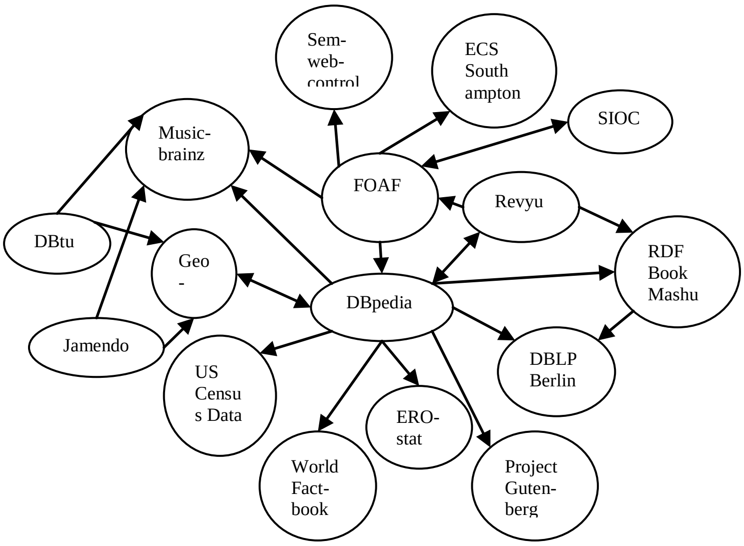

Figure 5. Linked data sources

6. RELATED WORK

The free encyclopedia Wikipedia has been tremendously successful due to the ease of collaboration of its users over the Internet. The Wikipedia wiki is the representative of a new way of publishing and currently contains millions of articles. It is a natural idea to exploit this source of knowledge. In the area of Machine Learning the Information Retrieval community has applied question answering, clustering, categorization and structure mapping to Wikipedia content. There is a vast body of works related to the semantification of Wikipedia.

6.1 FOAF

A popular application of the semantic web is Friend of a Friend (or FOAF), which uses RDF to describe the relationships people have to other people and the "things" around them [8]. FOAF permits intelligent agents to make sense of the thousands of connections people have with each other, their jobs and the items important to their lives; connections that may or may not be enumerated in searches using traditional web search engines. Because the connections are so vast in number, human interpretation of the information may not be the best way of analyzing them. FOAF is an example of how the Semantic Web attempts to make use of the relationships within a social context.

6.2 KIWI: Knowledge in a Wiki

Knowledge management in software development and project management is an exciting problem, as it involves tacit knowledge (e.g. about processes), distributed knowledge (different people, different systems), and many different kinds of semantically rich content (e.g. source code, documentation, tutorials, project work plans) that is strongly connected on the conceptual level. Current knowledge management systems only insufficiently support knowledge management in such areas, as they are not flexible enough to handle and integrate these kinds of content and provide only insufficient support for tacit knowledge.

The objective of the project KIWI is to develop an advanced knowledge management system (the "KIWI system") based on a semantic wiki that will address this problem. This system will support collaborative knowledge creation and sharing, and use semantic descriptions and reasoning as a means to intelligently author, change and deliver content. A particularly salient aspect of combining wikis with advanced semantic technologies is that the wiki still is a generic and flexible tool, but semantic technologies allow providing specific support for the user based on domain, context, role, and experience.

6.3 Freebase Wikipedia Extraction (WEX)

Freebase is another interesting approach in semantic web. The project aims at building a huge online database which users can edit in a similar fashion as they edit Wikipedia articles today [9]. Freebase is a collaborative project and may be edited by anyone, but it doesn't run on media wiki software.

Freebase is a repository of structured data of almost 22 million entities. An entity is a single person, place, or thing. Freebase connects entities together as a graph. Freebase use Ids to uniquely identify entities anywhere on the web. Freebase uses query language for querying its data. This query language is MQL (Metaweb Query Language). Freebase defines its data structure as a set of nodes and a set of links that establish relationships between the nodes. Because its data structure is non-hierarchical, Freebase can model much more complex relationships between individual elements than a conventional database.

Freebase and DBpedia both extract structured data from Wikipedia and make RDF available. Both are part of web of data and there are many connections of topics between them. Freebase imports data from a wide variety of sources, not just Wikipedia, whereas DBpedia focuses on just Wikipedia data.

6.4 The Semantic MediaWiki project

Semantic MediaWiki (SMW) project aims at enabling the reuse of information within Wikis as well as at enhancing search and browse facilities. SMW is a semantic wiki engine that enables users to add semantic data to wiki pages [11]. This data can then be used for better searching, browsing, and exchanging of information. While traditional wikis contain only text which computers can neither understand nor evaluate, SMW adds semantic annotations that allow a wiki to function as a collaborative database. Semantic MediaWiki introduces some additional markup into the wiki-text which allows users to add "semantic annotations" to the wiki. While this first appears to make things more complex, it can also greatly simplify the structure of the wiki, help users to find more information in less time, and improve the overall quality and consistency of the wiki.

DBpedia and SMW both deal with structure data represented in RDF. DBpedia concentrates in transforming all Wikipedia pages to semantic web pages. SMW concentrates in adding semantics to newer pages.

7. CONCLUSIONS AND FUTURE WORK

As we outlined in the paper, we discuss the approach of extracting information from Wikipedia. This is concentrated on infobox template of Wikipedia page. Also, we discuss the output of some Wikipedia resources and the extractors we get. We clarify about integrating DBpedia with other sources of the web, and how to make the semantic web as a nucleolus for a web open data.

Till now, the extraction framework can't get the full vision of semantic web. Tim Berners-Lee originally expressed the vision of the Semantic Web as follows: I have a dream for the Web in which computers become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A ‘Semantic Web’, which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The ‘intelligent agents’ people have touted for ages will finally materialize.

All related work we discuss may not have completed vision. And this is the problem of semantic web. Also, we can’t guarantee that DBpedia project will be better than current Wikipedia. As future work, we will first concentrate on improving the quality of the DBpedia extraction process. Also, we plan to extract information from other Wikipedia templates to extract more data from Wikipedia resource. DBpedia is a major source of semantic web data, we hope interlinking DBpedia with further data sources. It can serve as a nucleus for emerging web of open data.

8. REFERENCES

[1] A. Jentzsch, “ DBpedia – Extracting structured data from Wikipedia,” Presentation at Semantic Web in Bibliotheken, Cologne, Germany, November 2009.

[2] C. Bizer, “D2R MAP: A database to RDF mapping language,” In WWW, 2003.

[3] C. Becker, “ DBpedia – Extracting structured data from Wikipedia,” Presentation at Wikimania 2009, Buenos Aires, Argentinia, August 2009.

[4] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak and S. Hellmann. “DBpedia – A Crystallization Point for the Web of Data”, Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Issue 7, Pages 154–165, 2009.

[5] Christian Bizer, Richard Cyganiak and Tom Heath, “How to publish linked data on the web” http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial/, 2007.

[6] Eric Prud'hommeaux and Andy Seaborne, “ SPARQL Query Language for RDF”, http://www.w3.org/TR/rdf-sparql-query, January 2008.

[7] G. Antoniou and Frank van Harmelen, A semantic Web primer - The MIT Press, second edition, 2008.

[8] J. Golbeck, M. Rothstein, “ Linking Social Networks on the Web with FOAF,” A Semantic Web Case Study Proceedings of the Twenty-Third Conference on Artificial Intelligence AAAI, 2008, p.1138-1143.

[9] K. Bollacker, C. Evans, P. Paritosh, T. Sturge and Jamie Taylor, “ Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge,” Proceedings of ACM SIGMOD international conference on Management of data, 2008.

[10] M. Krötzsch, D. Vrandecic and Max Völkel, “Wikipedia and the Semantic Web - The Missing Links,” Proceedings of Wikimania, 2005.

[11] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller and Rudi Studer, “ Semantic Wikipedia,” 15th international conference on World Wide Web, 2006, pages 585-594.

[12] N. Shadbolt, W. Hall and Tim Berners-Lee, “ The Semantic Web Revisited,” IEEE Intelligent Systems, june-2006.

[13] R. Hahn, C. Bizer, C. Sahnwaldt, C. Herta, S. Robinson, M. Bürgle, H. Düwiger and U. Scheel, “Faceted Wikipedia Search”, 13th International Conference on Business Information Systems, Germany, May 2010.

[14] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, Z. Ives, R. Cyganiak, “ DBpedia: A Nucleus for a Web of Open Data,” 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, 2007.

[15] S, Auer and J. Lehmann, “What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content,” In Franconi et al. (eds), Proceedings of European Semantic Web Conference (ESWC’07), Springer, 2007, pp. 503–517.

[16] Tim Berners-Lee, “Linked dat,” http://www.w3.org/DesignIssues/LinkedData.html, 2006.

[17] Tim Berners-Lee, “ Primer: Getting into RDF & Semantic Web using N3,” http://www.w3.org/2000/10/swap/Primer.html, 2005.