Extraction of Biographical Data from Wikipedia

Robert Viseur
Centre of Excellence in Information and Communication Technologies, Rue des Frères Wright, 29/3, B-6041 Charleroi Belgium
Université de Mons, Faculté Polytechnique, Rue de Houdain, 9, 7000 Mons, Belgium

Keywords: Wikipedia, Dbpedia, Biography, Text Mining, Open Data.

Abstract: Using the content of Wikipedia articles is common in academic research. However the practicalities are rarely analysed. Our research focuses on extracting biographical information about personalities from Belgium. Our research is divided into three sections. The first section describes the state of the art for data extraction from Wikipedia. A second section presents the case study about data extraction for biographies of Belgian personalities. Different solutions are discussed and the solution adopted is implemented. In the third section, the quality of the extraction is discussed. Practical recommendations for researchers wishing to use Wikipedia are also proposed on the basis of our case study.

1 INTRODUCTION

Wikipedia (wikipedia.org) is a collaborative multilingual encyclopedia launched in 2001. The project has been supported financially since 2003 by the Wikimedia Foundation (wikimediafoundation.org). The volume of the encyclopedia has grown steadily since its inception. In January 2013, the largest editions of Wikipedia were the English edition (more than four million items), the German edition (more than one and a half million items), the French edition (more than one million three hundred thousand items) and the Dutch edition (over one million one hundred thousand items).

In recent years, academic research and practical examples of using Wikipedia content have increased. Hu et al. (2009) used it to improve the performance of a system for clustering documents. Kazama et al. (2007) and Charton et al. (2010) used it to improve a named entity recognition system. Buscaldi and Rosso (2006) improved the performance of a Question Answering technology. The BBC used it to be able to make the interconnection of information in its internal databases, and the enrichment by external data sources (Kobilarov et al., 2009). The "Exploiting Wikipedia" query on the scientific search engine Google Scholar (scholar.google.fr) returns more than 22,000 results!

Our research relates to the extraction of biographical data about people from Belgium. Using Wikipedia to supply a biographical database seems appropriate, due to the breakdown by type of content within the encyclopedia. Indeed the articles related to biographies represented 15% of the total content in January 2008, behind the articles about culture and the arts (Kittur et al., 2009).

However, several questions arise.

a) The French, German and Dutch editions of Wikipedia are useful, because these languages are the three national languages of Belgium. However it is difficult, on this basis, to identify the volume of content about Belgium rather than France, Germany or the Netherlands.
b) Many papers exploit Wikipedia content. However, few give guidance concerning the practical difficulties associated with the extraction of data from Wikipedia. Successful extraction involves knowing how to identify relevant articles but also to be able to extract the desired data from the content of articles.

Our research is organized into three sections. The first section will provide a state of the art about data extraction in Wikipedia. A second section will present the case study of the extraction of biographical data about Belgians. Different solutions will be discussed and the chosen solution will be implemented. In the third section, the quality of the extraction will be discussed. Practical recommendations for researchers wishing to use

Viseur R..
Extraction of Biographical Data from Wikipedia.
DOI: 10.5220/0004595302480252
In *Proceedings of the 2nd International Conference on Data Technologies and Applications* (DATA-2013), pages 248-252
ISBN: 978-989-8565-67-9

Wikipedia content will also be offered on the basis of our practical example.

2 STATE OF THE ART

The extraction of biographies was already done by Biadsy et al. (2008). However the approach adopted by the authors was different from ours. They developed a system of multi-document summaries, based on a classifier of biographical sentences and on a scheduling component for sentences deemed of interest. They were based on the articles using the Wikipedia template for biographies and were able to extract nearly 17,000 articles. The treatment was made on the Wikipedia XML copy available online.

In practice, the use of XML copies is not the only way to manipulate the contents of the encyclopedia. On the one hand, information extraction is possible using reverse engineering tools directly on the pages published online. On the other hand, a structured version of Wikipedia has been available since 2007, called Dbpedia.

DBpedia (dbpedia.org) is a community effort that started in 2007 (Auer et al., 2007). It aims to extract structured information from Wikipedia and to make this information available on the Web. The extraction process is based on copies of the Wikipedia database ("database dump"). The data is updated through the use of flow referencing updates of the encyclopedia (Hellmann et al., 2009). The extractor is based on the content of articles, and especially on the associated Infobox. The Infoboxes appear in tabular form in the upper right-hand corner of numerous articles and present factual information.

The content extracted from the encyclopedia is converted into RDF format. Several mechanisms are suggested to access and explore DBpedia: access to RDF data by URI (Universal Resource Identifier), use of Web agents (e.g. browsers for the semantic Web) and SPARQL access points to query DBpedia using language referring to the SQL used for relational databases.

DBpedia appears as a partial solution for the extraction of data from Wikipedia content. The interrogation facility permitted by the SPARQL query language for the identification of relevant articles makes it an attractive tool. However, DBpedia has several limitations.

Firstly, the language coverage of DBPedia is currently limited to 13 languages (see "International DBpedia chapters", dbpedia.org). At its inception in 2007, DBpedia was only available in English. A project for the French language was launched in late 2012. Called Sémanticpédia (www.semanticpedia .org) it combines the efforts of the French Ministry of Culture and Communication, Wikimedia France and INRIA to produce a French version of DBpedia (fr.dbpedia.org).

Secondly, the extraction process is based primarily on the content of Infoboxes (Auer et al., 2007); (Hellmann et al., 2009). However, a quick review of Wikipedia articles shows that not all the pages of the encyclopedia offer an Infobox, and that they are not always complete. Part of the information contained in the articles thus escapes from the extractors. However DBpedia already claimed nearly 2 million references at its inception (Auer et al., 2007).

3 CASE STUDY: EXTRACTING BIOGRAPHICAL DATA ABOUT BELGIANS

3.1 Identification of Relevant Articles

We first compared two approaches: firstly, the querying of DBpedia from English and French access points and, secondly, the identification of relevant articles using techniques of crawl on the website of the encyclopedia.

The querying of English and French DBpedia was performed with the SPARQL query language, by using the "birthplace" property (i.e. "Belgique" for the French language and "Belgium" for the English language).

The identification of Belgian personalities' biographies was performed in two stages. The first step takes as its starting point the Wikipedia page about Belgians (http://fr.wikipedia.org/wiki/ Cat%C3% A9gorie:Personnalit%C3%A9_belge), starting from the Belgian Wikipedia portal (http://fr.wikipedia.org/wiki/Portal:Belgium). A recursive crawl was processed on this page and the pages of the following categories in order to identify the category pages containing information about Belgians. This mechanism allowed us to find more than 700 relevant categories. The URLs of these categories were stored. The second step then explored the category pages and identified Wikipedia articles devoted to Belgians. The URLs of these files were saved in a file. More than 10,000 items were collected through this method (see Table 1).

The volume of the classical method by crawl of Wikipedia rather than querying DBpedia proves so much more fruitful.

Table 1: Number of items per method.

	Number of results
DBpedia(en)	899
DBpedia (fr)	200
Wikipedia (fr)	10,884

3.2 Data Extraction from the Text

3.2.1 Extraction Process

A copy of the articles was saved locally. In practice, we worked on the text of the articles in the specific Wikipedia format. This version is accessible from URLs for which the template is http://fr.wikipedia.org/w/index.php?action=raw&title=xxxxx, and provides a plain text (text + Mediawiki syntax) without HTML tags and starting with an Infobox when there is one.

The plain text is analysed through two operations. The first one is to extract the Infobox when it exists. The second one is to identify sentences in the biography that may contain important biographical information such as date of birth, date of death and professional activity. In practice, the first sentence of the article is always used, because it often contains by convention the most important information about the person. It may be supplemented by a second sentence, if it matches with a set of triggering words. This treatment results in a condensed biography, which is saved for each article. These condensed biographies then pass through a set of regular expressions to extract the date of birth, the date of death (if the person is dead) and his/her profession. This structured data is stored in a CSV file.

This file contains 10,610 entries, with the following fields: name, date of birth, date of death, professional activity, URL of the category and URL of the article in HTML format (see Table 2). From an initial total of 10,884 items, 57.6% allow extraction of the date of birth, 26.9%, date of death and 56.3%, professional generally provides an alternative information if the extraction failed (categories often indicate a profession or a social function). Only 27.4% of the articles have an Infobox.

Table 2: Volumetrics (extraction process).

Number of articles	10,884	100.0%
Number of Infoboxes	2,980	27.4%
Numbered of condensed biographies	10,610	97.5%
Number of successful extractions		
Date of birth	6,269	57.6%
Date of death	2,936	26.9%
Profession	6,129	56.3%

3.2.2 Main Difficulties

We met four main difficulties.

Firstly, the items are accompanied by an Infobox in less than one out of three cases. This makes it necessary to use text analysis techniques to achieve the extraction of dates (birth, death) and professions . The extraction of dates is particularly difficult because the articles often include other dates (dates related to important events in the people's lives). The extraction uses a set of regular expressions, which present writing difficulties for the non-specialist.

Secondly, even when an Infobox is present, the field names of the Infobox are not homogeneous. The date of birth is announced by *date_naissance*, *date naissance*, *date de naissance*, *date_de_naissance* or *naissance*. A preliminary grouping is necessary. This presents no big technical difficulty.

Table 3: Heterogeneity of date formats (examples).

([[Bree]], [[12 avril]] [[1876]] – [[Ixelles]], [[14 septembre]] [[1953]])
([[Pétange]], {{Date de naissance|12|juillet|1817}} - Pétange, {{Date de décès|14|mai|1898}}])
né le [[12 janvier]] [[1597]] à [[Bruxelles]] ([[Belgique]]) et mort le [[12 juillet]] [[1643]] à [[Livourne]] ([[Italie]])
'''Ellen Petri''' (née le 25 mai [[1982]], [[Merksem]] ([[Anvers]]))
'''Paul Deschanel''', né le {{date|13|février|1855}} à [[Schaerbeek]] ([[Bruxelles]]) et décédé le {{date|28|avril|1922}} à [[Paris]]
'''Robert Gruslin''' né à [[Rochefort (Belgique)|Rochefort]] le [[18 mars]] [[1901]], décédé à [[Profondeville]] le {{1er juin}} [[1985]]

Thirdly, the date formats are not homogeneous, either in the text or in the Infobox (see Table 3). Dates can be written with numbers only, with the month in letters or be supplemented by other information such as place of birth or the type of activity for which the person is famous.

Fourthly, the screening of sentences useful for data extraction requires a more advanced implementation than the technique used here. A classifier as implemented by Biadsy et al. (2008) deserves an investment to improve the overall performance of the extraction.

3.2.3 Error Rates

The evaluation was conducted on a set of 2,980 entries (i.e. entries including an Infobox). The dates of birth extracted from the text of articles were compared with those provided in the Infobox. The content of the infobox is structured. The extraction is significantly simplified, and the data extracted can be considered as free of extraction errors.

Table 4: Extraction Error Rate (Date of Birth).

Total number of items	2,980	100%	
No possible comparison	1,336	44.8%	
Number of Info-boxes without date	743	24.9%	
Possible comparison	1,644	55.2%	100%
Identical dates	1,486		90.4%
Different dates	158		9.6%
Partial information	126		7.7%
Extraction error	32		1.9%

A comparison was made between the data extracted from the text of Wikipedia articles and data extracted from the Infobox (see Table 4). The test was carried out for 2,980 birthdates (100%). The comparison was performed on 1,644 dates for which the data was present in the Infobox and in the result of the extraction from the text of the article. Different dates are found in 9.6% of cases. However, 7.7% of the dates were correct but had incomplete information. Typically, the year of birth was extracted, but not the full date (eg *mai 1988* vs. *1988*). The information extracted from the text could be more complete than that extracted from the Infobox. The information could be found in the text and not in the Infobox.

The presence of information in the Infobox and not in the text is due to extraction errors. In practice, the information in the Infobox always seems to be present in the text. This finding provides a lower limit to the rate of failed extractions from the text of 19.9%.

Almost two thirds of people are born after 1900 (63.1% of dates of birth given in the Infoboxes). The low number of dates of death would be due to the average age of registered persons rather than extraction errors.

This method presents two difficulties. On the one hand, date formats may differ between data extracted from the text and data extracted from its Infobox (example: *8 mars 1965* vs. *8 03 1965*). A method for converting dates is therefore necessary to standardize the format. Mediawiki tags and additional information can also accompany the date of birth (e.g. *date_de_naissance = [[28 juillet en sport|28 juillet]] [[1982 en football|1982]]*). On the other hand, the structure of the Infobox is not standardized and field names may vary from one item to another.

4 DISCUSSIONS AND PERSPECTIVES

This extraction work was initiated with the thought that the use of DBpedia would easily allow us to get biographical data we wanted with the SPARQL query language. A first test showed that the volume available with DBpedia was significantly less than that which could be obtained from conventional techniques of crawling the Wikipedia website. The DBpedia project is essential for researchers participating in projects related to linked data or wishing to have a controlled vocabulary. However, it shows its limits in terms of completeness on specific topics.

The existence of the DBPedia project and the visibility of a structured Infobox may give the impression that Wikipedia lends itself to easy data retrieval. However, it is clear from our experiments that, firstly, the Infoboxes are far from systematic (less than 30% of the articles considered possess one) and, on the other hand, the structure of the Infobox is not completely homogeneous. However the existence of a set of agreements in the form of markup or the turns of sentences, in terms of dates or professions, makes it feasible to extract content from articles without requiring the use of sophisticated techniques.

This research offers several perspectives. Firstly the influence of the formulation of requests on SPARQL results should be studied further. Secondly the consistency of the information extracted in different languages should be checked. Thirdly a comparison with more general extraction methods and tools (e.g. OpenNLP, ReVerb or TextRunner) should be processed. Fourthly the reliability of data in the encyclopedia should be checked. This work is ongoing and is based on a comparison with reference data. Disambiguation is one of the challenges to be addressed in order to automate this comparison.

REFERENCES

Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., Ives Z., 2007. DBpedia: A Nucleus for a Web of Open Data, *Lecture Notes in Computer Science*, Vol. 4825, pp 722-735.

Bekavac B., Tadić M., 2008. A Generic Method for Multi Word Extraction from Wikipedia, *Proceedings of the Int. Conf. on Information Technology Interfaces*, June 23-26, 2008.

Biadsy F., Hirschberg J., Filatova E., 2008. An Unsupervised Approach to Biography Production using Wikipedia, *Proceedings of ACL-08: HLT*, pp. 807–815.

Buscaldi D., Rosso P., 2006. Mining Knowledge from Wikipedia for the Question Answering task, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Charton E. Gagnon M., Ozell B., 2010. Extension d'un système d'étiquetage d'entités nommées en étiqueteur sémantique, *TALN 2010*, 19–23 juillet 2010.

Hellmann S., Stadler C., Lehmann L., Auer S., 2009. DBpedia Live Extraction, *Lecture Notes in Computer Science*, Vol. 5871, pp 1209-1223.

Hu X., Zhang X., Lu C., Park, E. K., Zhou, X., 2009. Exploiting Wikipedia as external knowledge for document clustering, *KDD '09 Proceedings of the 15th international conference on Knowledge discovery and data mining*.

Kazama J., Torisawa K., 2007. Exploiting Wikipedia as External Knowledge for Named Entity Recognition, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, June 2007, pp. 698–707.

Kittur A., Chi E.H., Suh B., 2009. What's in Wikipedia?: Mapping Topics and Conflict using Socially Annotated Category Structure, *Proceedings of the 27th international Conference on Human Factors in Computing Systems*, April 04-09, 2009.

Kobilarov G., Scott T., Raimond Y., Oliver S., Sizemore C., Smethurst M., Bizer C., Lee R., 2009. Media meets Semantic Web - How the BBC uses DBpedia and Linked Data to make Connection, *ESWC 2009*, pp. 723-737.