Exploiting Wikipedia as a Knowledge Base: Towards and Ontology of Movies

Rodrigo Alarcón, Octavio Sánchez, Víctor Mijangos

Grupo de Ingeniería Lingüística, Universidad Nacional Autónoma de México
Basamento de la Torre de Ingeniería, Ciudad Universitaria, México, D.F.
{ralarconm,osanchezv,vmijangosc}@iingen.unam.mx

Abstract. Wikipedia is a huge knowledge base growing every day due to the contribution of people all around the world. Some part of the information of each article is kept in a special, consistently and formatted table called *infobox*. In this article, we analyze the Wikipedia infoboxes of movies articles; we describe some of the problems that can make extracting information from these tables a difficult task. We also present a methodology to automatically extract information that could be useful towards the building of an ontology of movies from Wikipedia in Spanish.

Keywords: Wikipedia mining, Wikipedia infobox, ontology construction, semantic relation extraction, information extraction, natural language processing.

1 Introduction

Wikipedia is a free encyclopedia of open content that has become an important resource towards the construction of the Semantic Web. Since it beginnings, in the year 2001, the English version has achieve more than 2 million of articles, while the Spanish version has around 480 thousand of articles. All of the content has been written and edited by volunteers from different countries in many different languages, and it is covered by GFDL (GNU Free Document License), which makes possible to freely use them.

One important thing about the structure of Wikipedia is the social control executed by the community, which is able to avoid the spam, the nonsense and other kind of vandalism that is recurrent on some media sites. Besides, this same control makes possible to constantly increase the quality and precision of the articles.

Inside Wikipedia, there is an entry called *Wikipedia: Wikipedia in academic studies*[1], where it is possible to see the growth of academic interest in this encyclopedia. This interest is related to the use of Wikipedia on different academic studies and as a knowledge base for developing specific tools. On one hand, to mention a few, some works have focused on the social theme that represents Wikipedia [1] [2], some other have denounced inherent problems presented on this

[1] http://en.wikipedia.org/wiki/Academic_Research_on_Wikipedia.

kind of media sites [3], and others have obtained specific information and statistic data about the users [4]. On the other hand, Wikipedia has become a useful resource for the extraction of definitions, name entity recognition, machine translation or semantic relation extraction [5]. In this last field, Wikipedia represents a huge knowledge base that has made possible the developing of specific ontologies for the construction of the Semantic Web.

In this paper we present a work in process for the elaboration of an ontology of movies from Wikipedia on Spanish language. First we will briefly present an overview of some studies related to the use of Wikipedia for semantic relation extraction and ontology construction (2). Then we will explain the first step towards the elaboration of an ontology of movies (3). This step includes: a) the description of the so-called *infobox*, which is part of each movie of Wikipedia and contains specific data about the film (3.1); b) the specific relations to automatically extract (3.2); c) and our proposed XML schema to represent these relations (3.3). Finally, we will discuss our preliminary results and present the future work (4).

2 Wikipedia as a Semantic Knowledge Base

There is a growing interest of efforts to mine the information in Wikipedia for different purposes. As we have mentioned before, one of this interest is the extraction of semantic information that could be helpful on the process of *giving more meaning* to the Web. In Wikipedia, the meaning could be seen as the knowledge about things represented in different ways: definitions, descriptions, images, numeric data, etc. Furthermore, the meaning of each concept explained on the encyclopedia is related to the meaning of other concepts, which becomes a helpful semantic network to understand concepts on the field where they belong.

In this sense, Wikipedia represents a valuable source of knowledge to extract semantic information between concepts. A general overview of how Wikipedia could be used to extract concepts, relations, facts and descriptions can be found in [6]. Here, the authors explain the use of Wikipedia for natural language processing, information extraction and ontology building.

In [7], the authors describe a methodology that uses the links between categories to mine specific relations. They analyze some measures to infer relations and try to provide a semantic scheme in order to improve the search capabilities and to give the users meaningful suggestions to edit articles. In the same context, in [8] the authors use Wikipedia to develop a methodology for the automatic annotation of different semantic relations. This work is based on discovering lexical patterns that can be used to recognized specific relations between concepts. They evaluate the methodology by using a corpus and searching on it the relations founded in Wikipedia. Their results show that this kind of methodology could be a good starting point for automatic ontology construction.

The research presented in [9] shows how hyperlinked pages are used to generate a domain hierarchy by means of ranking articles that are strongly linked. These articles become a domain corpus for the automatic construction of an ontology. The same goal of obtaining ontologies through Wikipedia is described in [10], where the authors apply machine learning techniques to improve the performance of a system that mines the *infoboxes*. Finally, in [11] we can found another example of the use of Wikipedia for ontology construction, specifically for document classification.

This is not, and does not pretend to be an extensive list of all the works made about semantic relation extraction or ontology construction from Wikipedia. Our main purpose is to state both the interest that has woken up in the area of extraction and organization of semantic information, and some of the automatic analyzes and procedures that are possible to develop taking into account Wikipedia's structure. Nevertheless, as we will see in this paper, this structure is often not well organized and makes it difficult to implement automatic processes.

3 Towards an Ontology of Movies

In order to develop an ontology of movies we have stated three main steps that can lead us to our purpose. The first one is to collect our input corpus from Wikipedia movies articles and the analysis of the infobox structure on them. After that, the second step is the delimitation and automatic extraction of specific semantic information. Finally, as a third step we consider the implementation of the extracted information into a XML schema that will conform the basis for another later annotation schema.

3.1 Movies infobox structure

The first step of our methodology was to conform a corpus from the articles of the *films by year* category. We use the *categories tree* option to find a list of the movies titles from the year 1892 to 2008[2]. After that, we use the *export pages* option to retrieve all the articles of this list. We found a total of 5,561 articles, where the opening and closing infoboxes tags ({{Fields…}}) was on 5,092 cases. This late number represents the total of articles from our corpus.

After that, we analyze the *infobox* of each entry. The *infobox* is a resource used on Wikipedia to summarize and group the information about specific data on some articles. In general words, its purpose is to make the information on a more available format and it can be use as a resource to other applications.

In Spanish language, there are 49 proposed fields for the infobox, where only two are consider as required: *film title* and *original title*. The infobox will be framed in {{Fields…}}, and each field inside will be preceded by a vertical bar "|" and followed by an equal sign "=" and the specific information. Fields without descriptions will remain empty after the equal sign. That means it will have the following structure:

```
| Field = description of the field
```

An example could be the next one:

```
| genre = Science fiction
```

[2] Data was collected on February 2009.

The whole fields used in the movies infoboxes from Wikipedia in Spanish can be found in table 1.

Table 1. Infobox template in Spanish.

Fields		
título original	diseño de producción	duración
título	guión	clasificación
índice	música	idioma
imagen	sonido	idioma2
nombre imagen	edición	idioma3
dirección	fotografía	idioma4
dirección2	montaje	productora
dirección3	vestuario	distribución
dirección4	efectos	presupuesto
dirección5	reparto	recaudación
dirección6	país	precedida_por
dirección7	país2	sucedida_por
dirección8	país3	imdb
dirección9	país4	filmaffinity
ayudantedirección	estreno	sincat
dirección artistica	estreno1	
producción	género	

From the table above we can see the different kind of information that the fields can introduce. We see information about *dirección* (direction), *estreno* (premiere), *idioma* (language, language2, language3, etc.), as well as *país* (country country2, country3, etc.), IMDb (Internet Movie Data Base) or Filmaffinity links (external Web sites with movies information).

The 49 fields from this table are the suggested ones in the official Wikipedia movies infobox template. Nevertheless, in our corpus we found several empty fields. We automatically found a total of 94,584 fields occurrences, while 30,742 cases where empty (32.48% of the whole occurrences).

Furthermore, one of the problems presented in the infoboxes is the lack of standardization. Some of the elements established by Wikipedia are written in an indistinctive way by the authors of the articles, while others have typographical errors. For example, the field *dirección* (direction) appears also as *director* (director); the field *título original* (original title) can be found as *título en España* (title in Spain), *título principal* (main title), *título traducido* (translated title), among others. More complicated is the case of *estreno* (premiere), which presents variations like *año* (year), *fecha* (date), *fecha de estreno* (premiere date), or *primera emisión* (first emission).

Typos are another common lack of standardization. For the field *género* (genre) we can find mistakes like **gènero*, **genero* or **genro*.

In the corpus we can also find the case of another fields that are not proposed in the original schema, such as *asistente de artes marciales* (martial arts assistant), *calificación* (qualification), *premios* (awards), *Myspace*, and so on. In this case we found a total of 205 non-official fields.

If we compare the schema in Spanish to the English one, we can notice that the latter infobox contains fewer fields, which probably allows to be more standardized at the moment to put it into practice. The fields of the movies infobox in English can be seen in table 2.

Table 2. Infobox template in English

Fields		
name	starring	country
image	music	language
image_size	cinematography	budget
caption	editing	gross
director	studio	preceded_by
producer	distributor	followed_by
writer	released	
narrator	runtime	

Here we can observe a total of 22 fields, comparing to the 49 in the Spanish template. It is important to notice the fact that most of another languages follow a similar structure like the one described for English. There is a similar template to the English movies infobox in French Wikipedia, with some added elements like format, awards, and *IMDb*. In Italian, the infobox determines general fields for different genres of films: *generic*, *animation* or *film a episodi* (films conformed from several short films), with specific fields for each genre; while in German, the fields specifies a more generic data, i.e., *title*, *original title*, *producer or cameraman*.

In infoboxes of different languages, the most common fields are *title*, *director* and *premiere*. There are also coincidences in other fields, for example *music* and *photography*. Between English and Spanish there is a coincidence in *preceded_by* and *followed_by*. Furthermore, in Spanish, as well as in French, there is the field of *IMDb*, while Italian or English do not include. However, in English links to *IMDb* or *Allmovie* can appear within the article as external links and not inside the template of the infobox. These external links are also a valuable information to extend the semantic data for an ontology, as they can add more information about the films that does not appears in Wikipedia, or be used to complete the empty fields of the infoboxes. Nevertheless, there is also no consistence between the occurrences of the tags with external links. In our corpus, the *IMDb* tag occurs approximately in the 80% of the articles, while *Filmaffinity* occurs around in the 5%.

3.2 Extracting specific relations data

Theoretically, the structure of the infoboxes contains information that should be exploited with relative easiness. We decide to automatically extract the *title*, *original title, director*, *premiere year* and *genre*, in order to generate a database with all of this information. Although, not all of this information is present on all the movies articles founded in the *films by year* category.

As we have mentioned before, there are some inconsistencies within the name of the fields, their completeness, or the way the authors write them. In the case of *director* field, we found it with complete information in the 5,092 occurrences of the articles with infoboxes, however the field *genre* occurs only in 4,499 of these cases. Taking into account that the inconsistencies of the metadata make more difficult the process of automatic relation extraction from the films information, we achieve to obtain the data through the process we hereby describe.

From our corpus, we find out that 5,092 articles contained at least one director, although the field name from many of them was not the same and a review had to be made in order to compile a list of *ad hoc* synonyms for searching this specific field. The synset was formed by *dirección* (direction), *director* (director) and *dirigida* (directed). Also, after the equal sign that should follow the name of the field, the kind of following blanks was not always the same. Sometimes there were tabs; some other, more than one simple space; and even other, without spaces. Many of the director's names are also entries of Wikipedia, so many users decided to establish links to their names, using the symbol "[[" followed by the name of the director and closing with "]]". This has the purpose of specifying to the wikiengine that there is a link: [[link to the article]]. But not all of them had those brackets, and it caused troubles while parsing the data with the aim of recovering the director's name of the film associated with the title of the entry.

The same problems were founded when we tried to mine the original title of the movie. Despite the fact that this field does appear in all the infoboxes, not all of them appear with information, which means that there are articles with the *original title* field empty. It does not contain information in 195 articles occurrences in the corpus.

With the *premiere* field it was also problematic to extract the information, because most of the films had different words to express the premiere year, for example *año* (year), *fecha de estreno* (premiere date) or **añoacceso* (acces year). In this case we decide to mine only the *año* (year) and *estreno* (premiere) variants, because of the wide range of structural possibilities. We found that 23 films infoboxes do not contain a premiere year, sometimes it was in the title and sometimes were completely absent.

Other field we exploited was the one of *género* (genre), which also present some inconsistencies that could be attached to human errors at the time of transcribing the template. This field was empty on 593 occurrences in our corpus and is the more unused one.

Summarizing, we can find the number of occurrences for each field in table 3:

Table 3. Numerical data found over the analysis of infoboxes

Field name	Number of full fields	Number of empty fields
Director	5,092	0
Título	5,092	0
Título ID	5,092	0
Título orignal	4,897	195
Año	5,069	23
Género	4,499	593
Director	5,092	0
Título	5,092	0

From the table above we can see the three fields with empty information: *premiere* or *year*, *original title* and *genre*. The first one was empty only in 23 articles, while the last one in more tan 500 cases. It is important to mention that the title of the movies was not obtained from the infobox but directly from the XML given by the Wikipedia, mainly because it is well demarcated by the labels <*title*> </*title*>; in the same way, we obtained the id used by the Wikipedia to identify each article.

Despite the inconsistencies and typos that make difficult the automatic process, in 4,499 cases all the information that we were trying to mine was complete. We consider that this number represents a good starting point to conform the basis of a first schema that could be later extended.

3.3 Proposed XML schema

With the data from the infoboxes that were exploited, we decided to generate a first XML scheme, which should give basic information about the film. This scheme can be expanded as we extend our extraction processes of the information contained in the Wikipedia articles.

To make this scheme, we decided to take *director* field as the root XML tag. The first tag will consist of the director's name. Taking into account that directors can have more than one film, we decided to introduce a *filmography* tag to include them. This last tag will include each *film* with *title*, *original title, year* and *genre* tags. On the opening *film* tag we added an attribute with Wikipedia's title id number. An example of the schema can be seen below.

Proposed XML schema for the organization of movies data on Spanish Wikipedia.

```
<director>
    <name>Brian de Palma</name>
    <filmography>
        <film wiki_id="2022905">
            <title>Carrie: la ira</title>
            <original_title>The Rage: Carrie 2 </original_title>
            <year>1999</year>
            <genre>Terror </genre>
        </film>
        <film wiki_id="587196">
            <title>La Dalia Negra</title>
            <original_title>The Black Dahlia</original_title>
            <year>2006</year>
            <genre>Crimen, Misterio, Thriller</genre>
        </film>
        …
        …
        …
    </filmography>
  </director>
```

As we can see in this example, the root tag is <*director*></*director*>. It is followed by the director's name tag <*name*></ *name* >. At the same level there is the tag *<filmography></ filmography>*. This tag nests the film tag *<film wiki_id=“”></film>*, which contains the relevant information of each film: *<title></title>*, *<original_title></original_title>*, *<year></year>* and *<genre> </genre>*.

Based on the XML scheme, relational databases can be generated to manipulate the information that we have considered at this first stage of the ontology construction. As we have said, this is not the full final scheme because as more data is extracted, the more can be added. This scheme is currently based on Wikipedia films articles in Spanish language, however it can be extended to fit another kind of relevant information, for example the country, external links (*IMDb*) or the id of directors or genre from Wikipedia. Furthermore, it will be possible to use this scheme in order to exploit Wikipedia in other languages, which could make possible to fill the empty fields in one language by relating them with the information on another languages, as well as to make multilingual queries.

4 Conclusions and future work

Nowadays, Wikipedia can be explored with the aim of obtaining information on different ways. The information added in a manual way by the users is generally well organized and semi-structured. Also, many entries from Wikipedia have infoboxes with summarized specific information about the theme treated in the article. We have mentioned that the structure of Wikipedia has made possible to exploit the information in order to extract semantic data. The extraction of semantic relations is one of the growing interests aiming to the construction of the Semantic Web.

Even so the structure of Wikipedia, we have noticed some specific problems on automatically exploiting it. To summarize a few, there are: a) the fact that the field's names are not respected; b) typos by human errors; c) lack of information; and d) differences on the infobox structure between languages. The latter should not be seen as a problem, however it would be advantageous to have standard fields on different languages.

Aiming to the standardization idea, it would be useful that the Wikipedia's process of writing or editing an article use a check-bot to confirm the information of the infoboxes templates. Thus, the fields not belonging to the template would be alerted, as well as typos on the field names. Furthermore, the same check-bot could be used to seek the existing fields looking for inconsistencies in the infoboxes or the whole articles.

The work that we have presented here is a first approach towards the elaboration of an ontology of movies from the Wikipedia in Spanish. We have showed the kind of semantic relations that are possible to mine, as well as a first scheme to represent them. We are conscious that this scheme may well be improved for achieving a complete ontology of movies. The future work will include: a) to define a scheme to represent subject, relation and predicates between the extracted information, for example a RDF scheme; b) to implement this new scheme for making the information available and sharing it with systems dedicated to the construction of the Semantic

Web; c) to develop a movie-ontology query system capable of retrieve the information on specific ways related to *directors*, *titles*, *genres* and *years* fields.

Acknowledgments

This research was made possible by the financial support of CONACYT (82050) and DGAPA-PAPIIT (IN403108). The authors wish to thank Sarahi Abrego Romero for the proofreading of this paper.

References

1. Kittur, A., Chi, E., Pendleton, B. A., Suh, B., Mytkowicz, T.: Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie. In: 25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007), ACM, New York (2007)
2. Suh, B., Chi, E. H., Pendleton, B. A., & Kittur, A.: Us vs. Them: Understanding Social Dynamics in Wikipedia with Revert Graph Visualizations. In: Visual Analytics Science and Technology. pp. 163-170, IEEE-Press, New York (2007)
3. Potthast, M.,Stein, B., Anderka, M.: Automatic Vandalism Detection in Wikipedia. In: 30th European Conference on IR Research, ECIR 2008, pp. 663-668, Glasgow (2008)
4. Stein, K., Hess, C.: Does it matter who contributes: a study on featured articles in the german wikipedia. In: Proceedings of the 18th conference on Hypertext and hypermedia, pp. 171-174, ACM, New York (2007).
5. 99 Wikipedia Sources Aiding the Semantic Web. AI3, http://www.mkbergman.com/?p=417
6. Medelyan, O., Milne, D., Legg, C., Witten, I.A.: Mining meaning from Wikipedia. Hamilton, (2008)
7. Chernov, S., Iofciu, T., Nejdl, W., Zhou, X.: Extracting Semantic Relationships between Wikipedia Categories. In: Proceedings of the 1st Workshop on Semantic Wikis - From Wiki to Semantics, ESWC2006, Budva (2006)
8. Ruiz-casado, M., Alfonseca, E., Castells, P.: From Wikipedia to Semantic Relationships: a Semi-automated Annotation Approach. In: Proceedings of the 1st Workshop on Semantic Wikis - From Wiki to Semantics, ESWC2006, Budva (2006)
9. Cui, G., Lu, Q., Li, W., Chen, Y.: Corpus Exploitation from Wikipedia for Ontology Construction. Conference on Language Resources and Evaluation, LREC2008, Morocco (2008)
10. Wu, F., Weld, D. S.: Automatically Refining the Wikipedia Infobox Ontology. In 17^{th} International World Wide Web Conference, Beijing (2008)
11. Kozlova, N.: Automatic Ontology Extraction for Document Classification. Ma. Thesis, Saarland University (2006)