An Experimental Approach for Collecting Snippets Describing the Relations between Wikipedia Articles

Andias Wira-Alam
GESIS – Leibniz Institute for the Social Sciences
Bonn, Germany
andias.wira-alam@gesis.org

Benjamin Zapilko
GESIS – Leibniz Institute for the Social Sciences
Bonn, Germany
benjamin.zapilko@gesis.org

Philipp Mayr
GESIS – Leibniz Institute for the Social Sciences
Bonn, Germany
philipp.mayr@gesis.org

ABSTRACT

In this paper, we deal with a simple scenario: a student Bob wants to know why "mathematics" is very important to "physics", or in a more specific case, why "differential equations" play a prominent role in the study of "fluid dynamics". In a different way, the scenario can also be stretched: Bob already had enough knowledge about "artificial intelligence" and now he wants to learn about "semantic web". According to the scenario, we try to examine the possibilities to give Bob an answer on the basis of the available knowledge resources on the Web.

We run a small experiment using Wikipedia articles as knowledge resources. First, we crawl a number of Wikipedia articles that are relevant to our scenarios and then we identify the *intra-wiki links* (links to the other Wikipedia articles) within the page body of the articles. For each link, we also identify the piece of text (snippet) where the link is located. Second, we apply a simple recursive algorithm to discover the relations (links and snippets) between articles, e.g. the possible connections between article "artificial intelligence" and "semantic web". Finally, we evaluate whether the found snippets could pithily describe the relations.

Keywords

Wikipedia, hyperlinks, experiment, text parsing, e-learning

1. BACKGROUND AND RELATED WORK

The Web is becoming the first place where people can get almost any information [2]. People also use the Web as source for their education or even just to satisfy their curiosity. Wikipedia and other Wikimedia Foundation's projects (e.g. Wikiversity) for example, they are slowly becoming a new form of knowledge resources: range from school students to senior scientists.

The semantic web technologies change the paradigm from *search engine* to *answering machine*. Based on [3], these technologies are considerably the answer to the above scenario. In this paper, nevertheless, we start with an hypothesis that a piece of text surrounding the links (snippet) might be a complement feature to the one of the important key elements, as described in [3], namely *typed links*. Furthermore, a collection of *snippets* might describe the relations between articles in more detail to the readers.

The DBPedia RelFinder [1] is a good example of finding the relations between Wikipedia articles based on the DBPedia Dataset. Despite the fact that finding the relations between articles is not new, we focus on the snippets collection and its evaluation rather than building rich ontologies. Another research effort, as described in [4], focuses on providing tools to efficiently access Wikipedia and Wiktionary, e.g. a visualization of the structure of a Wikipedia article. We see that [4] has no direct usage to our primary goal (scenario) as described in the abstract, although there are some overlaps.

2. HOW IT WORKS

A relation between two articles can be formally described as a simple triple (A, l_iB, s_{Al_iB}) where A is the source article, l_iB is the $i-$th intra-wiki link pointed to another article B, and s_{Al_iB} is snippet in A where l_iB is located. In this experiment however, we don't apply any particular algorithms to semantically identify s_{Al_iB}. We take the whole paragraph where l_iB is located in. Suppose B is the target article, so s_{Al_iB} might probably be one of the most significant snippets describing the relation between A and B.

In order to initially demonstrate the first proof of concept, we try to diligently find the relations of a pair: "artificial intelligence"[1] and "semantic web"[2] which both are represented as articles in Wikipedia[3]. In article "artificial intelligence" (AI), there is a link to "knowledge representation" (KR). In KR, there is a link to "semantic web". Based on those links, the snippets look as follows:

Knowledge representation and knowledge engineering are central to AI research.
Development of the **Semantic Web**, has included development of XML-based knowledge representation languages and standards, including RDF, RDF Schema, Topic Maps, DARPA Agent Markup Language (DAML), Ontology Inference Layer (OIL), and Web Ontology Language (OWL).

As alternate relations, in KR there is a link to "Web Ontology Language" (OWL), then to "ontology (information science)", then to "semantic web". Based on those links, the second snippets look as follows:

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.
.

[1] http://en.wikipedia.org/wiki/Artificial_intelligence
[2] http://en.wikipedia.org/wiki/Semantic_Web
[3] retrieved on March 19th, 2010.

Knowledge representation and knowledge engineering are central [...] Ontology Inference Layer (OIL), and **Web Ontology Language (OWL)**.
The Web Ontology Language (OWL) is a family of knowledge representation languages for authoring **ontologies**, and is endorsed by the World Wide Web Consortium. This family of languages is based on two (largely, but not entirely, compatible) semantics: OWL DL and OWL Lite semantics are based on Description Logic, which have attractive and well-understood computational properties, while OWL Full uses a semantic model intended to provide compatibility with RDF Schema.
Ontologies are used in artificial intelligence, the **Semantic Web**, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture as a form of knowledge representation about the world or some part of it.

Notice that the above snippets are collected by humans and also filtered to make sure that only the most relevant parts are displayed. Besides, they are also probably not the best or even there might be more other relations with better snippets.

3. EXPERIMENT SCENARIOS

We write a small program crawling Wikipedia articles (currently only English), given *source* and *target* articles, denoted as S and T respectively, and identify links as well as snippets. Each found relation is stored in the database as a simple triple as described above. We only identify links within the page body and we ignore all links point to non-Wikipedia articles. Besides, we don't consider "disambiguation pages" as well as links within nodes `<table>` and `<div class="rellink relarticle mainarticle">`, image captions, links under sections "See also" and "References". Briefly, we only consider as close as possible to the snippets expressed in natural language text.

The link distance, denoted as d, between two directly connected articles is 1 which means the minimum distance (d_{min}) that can be considered. Suppose that $\bar{n}$ is the average number of links within an article, the time complexity of the crawling in the worst case tends to be $O(\bar{n}^{d_{max}+1})$. For this experiment, therefore, we predefine d as $1 \leq d \leq 3$ due to limited time and resource. In consequence, there are at most 3 snippets describing each possible relation between S and T. In addition to this, there might also be no snippets in case that the minimum link distance between S and T is more than d_{max}.

To find all relations between S and T, we apply a simple recursive algorithm, namely *backtracking*. It starts with the T and find all its backlinks recursively until S has been reached. As a result, all found relations are shown as collections of snippets. In this experiment, we choose 8 pairs of articles, S and T, and find all snippets. For simplicity, we choose articles that are common knowledge and the average link distances likely close to d_{max}. But note that the relation between S and T is not reversible since the link is not bidirectional.

4. RESULTS AND EVALUATION

Table 1 shows the current results of our experiment[4]. n_{links} represents the number of the collected links given a *source* article with the maximum link distance d_{max}, n_c the number of snippets collections, $Source$ and $Target$ are titles (based on the article's URL), and $n_c(starred)$ denotes the number of n_c that we claim reasonable enough to describe the relations based on our domain expertise.

$Source$	$Target$	n_{links} / n_c / $n_c(starred)$
Bonn	Germany	92274 / 1232 / 12
Artificial_intelligence	Semantic_Web	49173 / 28 / 18
Differential_equation	Fluid_dynamics	41614 / 72 / 19
Information_theory	Information_retrieval	14961 / 14 / 5
Semantic_Web	Web_Science_Trust	17585 / 0 / 0
Mathematics	Physics	44894 / 885 / 139
Mathematics	Computer_science	44894 / 585 / 43
Philosophy	Politics	98686 / 301 / 93

Table 1: *source* and *target* articles.

According to the results, we propose that the number of snippets collections (n_c) indicates whether a particular article can be considered as "common knowledge" or "specific knowledge", although it has to be proven by further experiments. In addition, $n_c(starred)$ clearly indicates that our experiment has some merit for further development. On the one hand, as another interesting thing, n_c equal to 0 indicates that there is no relation, but on the other hand it might indicate that $Target$ is considered as "rare knowledge" in a particular area since it has no backlinks to $Source$.

As caveats, we realize that paragraph-based snippets are not effective enough and need huge effort to make it pithy. Moreover, our program is also not able to find "hidden links" which means that not all relations between articles are explicitly linked. As another fact, since program acts as a crawler, the contents of Wikipedia articles change unpredictably as well as the links.

5. FUTURE WORK

First, we see that our current work can be applied to support online learning as a learning resource tool. We plan to publish the results on the Web and let people evaluate the snippets collections by giving remarks, stars/ranking, notes, or even edit them if necessary. In addition, we state that the limitation on the maximum link distance is not necessary.

Second, the extraction of snippets is also an important issue to be improved in the first place. We consider to apply methods that can identify the most important or relevant parts of text to be taken as snippets. Natural Language Processing (NLP) is one method that could be employed for this purpose in the future.

6. REFERENCES

[1] DBPedia RelFinder. `http://relfinder.dbpedia.org/`.

[2] J. Hendler. Science and the semantic web. *Science Magazine*, 299(5606):520–521, 2003.

[3] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer. Semantic wikipedia. In *Proceedings of the 15th International Conference on WWW*, 2006.

[4] T. Zesch, C. Müller, and I. Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2008.

[4] retrieved on March 24-27th, 2010.