Twelve Years of Wikipedia Research

Judit Bar-Ilan
Department of Information Science, Bar-Ilan University
Ramat Gan, 5290002, Israel
Judit.Bar-Ilan@biu.ac.il

Noa Aharony
Department of Information Science, Bar-Ilan University
Ramat Gan, 5290002, Israel
Noa.Aharony@biu.ac.il

ABSTRACT
Wikipedia was formally launched in 2001, but the first research papers mentioning it appeared only in 2002. Since then it raised a huge amount of interest in the research community. At first mainly the content creation processes and the quality of the content were studied, but later on it was picked up as a valuable source for data mining and for testing. In this paper we present preliminary results that characterize the research done on and using Wikipedia since 2002.

Categories and Subject Descriptors
H.5.3 [**Information Systems**]: Group and Organization Interfaces – *Web-based interaction.*

H.3.5 [**Information Systems**]: Online Information Services – *Web-based services.*

General Terms
Measurement.

Keywords
Wikipedia, analysis, longitudinal trends

1. INTRODUCTION
Wikipedia is a unique, online, collaborative encyclopedia that was established in 2001 [5]. Since then it experienced exponential growth, and has been studied extensively. Among the studied topics related to Wikipedia are its structure, collaborative processes, reliability, content, improvements and research where Wikipedia data serve as input (e.g. data mining, semantics, and visualization). It is an attempt to create an online encyclopedia that presents the "wisdom of crowds" [1, 4]. As of December 2013 it contains 30 million articles written in more than 287 language editions [6] and more than 4.4 million articles in the English Wikipedia alone [7]. Wikipedia is one of the ten most visited sites on the web (see www.alexa.com).

This study aims to characterize research publications related to Wikipedia extracted from Elsevier's Scopus. Scopus is a multidisciplinary, citation database with extensive coverage. In particular we characterized how trends in studying Wikipedia during twelve years from 2001 to almost the end of 2013.

WebSci'14, June 23–26, 2014, Bloomington, IN, USA.
ACM 978-1-4503-2622-3/14/06.
http://dx.doi.org/10.1145/2615569.2615643

2. RESEARCH SETUP

2.1 Data Collection
Elsevier's Scopus (http://www.scopus.com) was searched on November 17, 2013. We searched for the occurrence of the term Wikipedia in the article title, abstract and keywords. Time span or article type were not limited. The number of retrieved records was 3582. Scopus is a multidisciplinary citation database. It was chosen over Thomson-Reuters' Web of Science (WOS, http://www.isiknowledge.com), because of its wider coverage of current publications especially of proceedings papers that constituted the majority of the retrieved documents from Scopus (2261 items, 63% of the total). The number of items retrieved from WOS was only 1,550, even though the proceedings citation databases were included in the search. Theoretically we could have used Google Scholar (GS, http://scholar.google.com), but on GS one can only search either in the text indexed by GS, which is often the full text of the article, or limit the search to title only. Looking at items that contain the term Wikipedia in the title only is too limiting (for example in the Scopus dataset only 864 out of the 3582 retrieved items contained the term Wikipedia in their title), and without any limitations the number of items reported by GS for the search Wikipedia was about 805,000 (this presumably includes papers that refer to a Wikipedia article for a definition). Thus it was not feasible to base our study on Google Scholar.

2.2 Content Categories and Reliability
The content of the items was analyzed. According to Krippendorff [2] content analysis is a "research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use" (p. 24).

The analysis was mainly based on the abstracts of the items. In case there was no abstract, or it was not possible to decide on the topic of the item based on the abstract, the full text of the item was consulted. We created a light-weight classification, which allowed us to classify the whole set of papers. Three facets were defined; the first described to which extent the item relates to Wikipedia (major, minor or unrelated). We encountered 614 cases of "unrelated" (17%) – although the Scopus records of these items included the term Wikipedia*,* but it was clear that the paper did not study or discuss. The second facet related to the actual topic of the item. Here we differentiated between articles that studied Wikipedia or its use (e.g. in education), and articles that used Wikipedia either as a source/resource for other research or used Wikipedia to test the feasibility and applicability of tools or methods developed for purposes not directly related to Wikipedia (e.g. the INEX initiative (https://inex.mmci.uni-saarland.de/) is using an xml collection based on Wikipedia to test the submitted outputs). The first category is called in what follows *about*, while the second is called *using*. The third facet concentrated on the item's approach: we explored if the item's focus was technological

or social/theoretical. We decided to include in the social/theoretical approach analyses and visualizations of Wikipedia. The technological approach for the *about* category only included tools developed for improving Wikipedia. We named the two categories in this facet *soc* and *tech* respectively.

The reliability of the categorization was assessed on a 10% random sample of the classified items by both authors [3, p.149]. The two coders agreed on 90% of the categorizations.

3. RESULTS

As mentioned before, 641 items were categorized as unrelated. The rest were either major (2301, 64%) or minor (667, 19%). From this point onward we only discuss the set of 2968 items that were categorized either as *major* or as *minor*.

In terms of topic, there were almost an equal number of items *about* Wikipedia (1431, 48%) as there were *using* Wikipedia (1537, 52%). As for approach, the *technological* approach was considerably more popular (1856 items, 63%) compared to the *social* approach (1112 items, 37%).

Figure 1 depicts the overall growth in the number of relevant publications indexed by Scopus (with *unrelated* excluded), and the growth in the topic and approach categories. We excluded 2013 from the graph, because we did not have the full data for that year. We see that the first papers using Wikipedia appeared in 2005, but since 2009 there are more papers that *use* Wikipedia than papers that are *about* Wikipedia. In terms of *social* versus *technological*, we see that at first the social aspects were emphasized, but since 2007 papers on technological aspects are much more frequent. Thus the crossover between *social* and *technological* occurred earlier than the crossover between *about* and *using*.

There is an overall growth in the total number of relevant papers published per year, but it seems that the number of publications per year plateaued and the growth rate is starting to level off. In order to support this finding we retrieved data from WOS, from Scopus and from the ACM Digital Library on May 7, 2014, assuming that the records for 2013 are complete by then. Figure 2 depicts the number of items per year in all three databases, with unrelated included in all years, since it was impossible to check the relatedness of the newly retrieved records by submission time. It clearly shows that the growth rate is slowing this down.

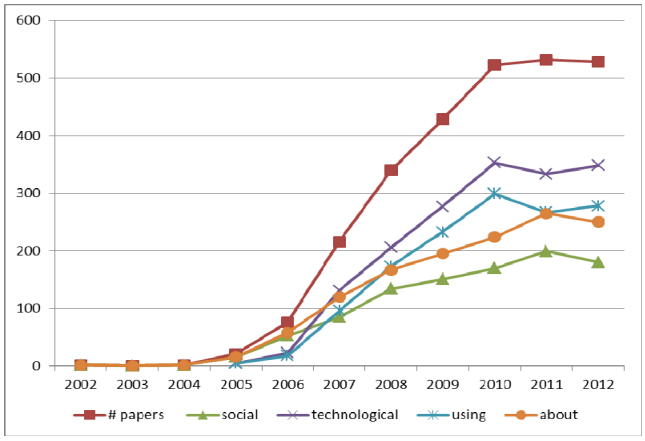

Figure 1: Number of papers per year, per topic and per approach

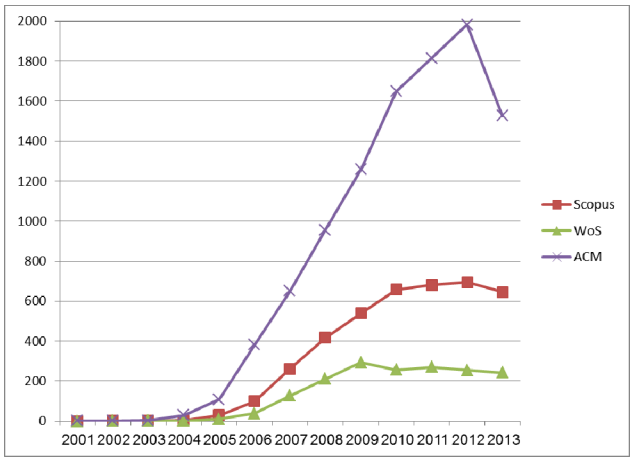

Figure 2: Number of retrieved items per year – Scopus, WOS and the ACM Digital Library

4. DISCUSSION AND CONCLUSIONS

We see that the growth rate is slowing down. This is somewhat surprising, since Wikipedia is an excellent, semi-structured, multilingual, interlinked and manually categorized data source that can and should be utilized extensively in NLP, IR, IE and ontology building. In addition, Wikipedia is a result of an unprecedented collaborative effort, and thus the social dynamics of editing, reaching consensus and creating quality encyclopedia articles should be of ongoing research interest.

This study is limited by the retrieval capabilities and the coverage of Scopus. We only considered records containing the term Wikipedia in the title, the abstract or the keywords, and thus might have missed some records. We could have possibly included the search term Wikipedians as well. The results highlight Wikipedia's importance in studying the Web and social media, and in advancing Web-based research.

5. REFERENCES

[1] Kittur, A., and Kraut, R. E. 2008. Harnessing the wisdom of crowds in Wikipedia: Quality through coordination. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work* CSCW 2008. New York: ACM Press, 37-36

[2] Krippendorff, K. 2013. *Content analysis: An introduction to its methodology*. Third Edition. Sage Publications.

[3] Neuendorf, K. A. 2002. *The content analysis guidebook*. Sage Publications.

[4] Surowiecki, J. 2004. *The Wisdom of Crowds*. Anchor Books. New York, NY.

[5] Wikipedia contributors. 2014. History of Wikipedia. http://en.wikipedia.org/w/index.php?title=History_of_Wikipedia&oldid=595511145

[6] Wikipedia contributors, 2014. Wikipedia. http://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=596604227

[7] Wikipedia: Statistics. 2013. http://en.wikipedia.org/w/index.php?title=Wikipedia:Statistics&oldid=587963650