Wikipedia in the Age of Siri: Task-based evaluation of Google, Wikipedia, and Wolfram Alpha

Hogyeong Jeong
Wisenut Research Lab
Seoul, Korea
hogyeong.jeong@gmail.com

WikiSym '12 Aug 27-29, 2012, Linz, Austria.
ACM 978-1-4503-1605-7/12/08.

Abstract

In this paper, we describe a task-based method to evaluate relative effectiveness of Wikipedia. We then use this method to compare Wikipedia against an internet search engine (Google) and an answer engine that uses structured data (Wolfram Alpha).

Author Keywords

Wikipedia, semantic knowledge, knowledge representation, search

ACM Classification Keywords

H.3.5 [Information Storage and Retrieval]: Online Information Services - *web-based services*; H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces - *collaborative computing*

General Terms

Measurement, Design, Experimentation

Introduction

Until recently, an encyclopedia provided the first and often the only resort for people looking up information. Then came internet search engines, which allowed free-form searches. More recently, sophisticated answer engines have been developed that exploit structured data. [2]

Whereas a search engine uses the whole accessible internet in answering queries, an answer engine such as Wolfram Alpha makes particular use of structured and hierarchical data to better understand the query and to better present the results. Meanwhile, Wikipedia represents a middle ground between the two, offering organized knowledge, which is sometimes accompanied by structured, albeit inconsistent, data (e.g. categories and infoboxes).

As Wolfram Alpha gathers more data, and Google incorporates more semantic information, awareness has been raised to allow more structured information into Wikipedia as well [4, 5]. Toolkits have also been developed to facilitate viewing and editing of structured information (e.g. categories or infoboxes) in Wikipedia [3, 1].

The focus of our research is to assess the relative merits of Google, Wikipedia, and Wolfram Alpha through a task-based evaluation. We believe that this research can establish a benchmark that can identify shortcomings in Wikipedia and provide guidance on how Wikipedia can evolve.

Experiment Setup

First, a set of fifteen factual questions were generated that could be answered using any of the three available options. These include questions such as "What was the magnitude of the major earthquake in Japan in 2011?" and "How many bytes are in a petabyte?".

A set of eight college-educated subjects answered these questions using Wikipedia and either Google or Wolfram Alpha; two subjects first used Wikipedia followed by Google (and two in reverse order) and two subjects first used Wikipedia followed by Wolfram Alpha (and two in reverse order). All subjects had at least some experience with Wikipedia and Google, but none had experience with Wolfram Alpha. All subjects reported being of at least average computer proficiency.

To prevent cases of using prior knowledge and guessing, the subjects were instructed to provide the web address of the page where they found the information. When subjects got stuck trying to find information, they were instructed to select "I can't find the information". Screencasting software was used to record the subjects' on-screen activities.

Hypotheses

Time spent in answering a question can be decomposed into the following components: question reading time, query input time, page loading time, and the answer selection time. In cases where the query does not bring about a desired page, additional query inputs/link clicks may be performed before selecting the answer.

The sum of the question reading time and the query input time is considered to be the **query generation time**. For cases where additional queries/link clicks were performed, the intermediate input and reading times were also added to the query generation time. Meanwhile, the sum of final reading time and the answer selection time is considered to be the **answering time**. For both times, the page load times were not considered.

Also, we define **coverage** as the percentage of the time that the subject reached the page required to answer the question, and define **answer rate** as the percentage of the time that the subject correctly answered the question.

There are two advantages that Wolfram Alpha has over Google and Wikipedia: first is in better interpreting the query, and the second is in better presenting the data. We expect quicker query generation times and greater coverage as a result of the former, and quicker answering times and higher answer rate as a result of the latter.

Results

Using the screencasts, we measured the **query generation** and **answering times**. **Coverage** was measured using the screencasts and the provided links, while **answer rate** was measured using the final answers. Results are below:

Method	**Query Gen**	**Answer**	**Cover**	**Ans**
Google	19.3s(5.2s)	17.6s(6.2s)	100%	93%
Wikipedia	21.2s(5.7s)	23.8s 6.5s)	100%	89%
Wolfram	13.2s(4.4s)	5.5s(2.6s)	100%	100%

Table 1: Comparative Evaluation Results (mean and standard deviation)

Because only factual questions were used for the study, Wolfram Alpha had an inherent advantage over the other systems, and this is well-demonstrated by the results. Subjects using Wolfram Alpha had significantly ($\alpha = 0.01$) faster query generation and answer times. Their answer rate was also perfect, compared to 93% and 89% for Google and Wikipedia respectively.

Conclusion and Future Directions

Our results show that having structured knowledge can indeed allow faster retrieval of information by allowing better interpretion of the query and better presentation of the results. This research confirms fears of others that Wikipedia should move towards having more structured underlying data. However, the questions used for this task were rather limited in scope. For the conclusion to be more valid, we imagine evaluating the systems using more varied sets of questions and tasks.

References

[1] Arnold, C., Fleming, T., Largent, D. L., and Lüer, C. Dynatable: a wiki extension for structured data. In *Int. Sym. Wikis* (2009).

[2] Berners-Lee, T., Hendler, J., and Lassila, O. The semantic web. *Scientific American 284*, 5 (2001), 34–43.

[3] Bostandjiev, S., O'Donovan, J., Hall, C., Gretarsson, B., and Höllerer, T. Wigipedia: A tool for improving structured data in wikipedia. In *ICSC* (2011), 328–335.

[4] Dohrn, H., and Riehle, D. Design and implementation of the sweble wikitext parser: unlocking the structured data of wikipedia. In *Int. Sym. Wikis* (2011), 72–81.

[5] Trattner, C., Hasani-Mavriqi, I., Helic, D., and Leitner, H. The austrian way of wiki(pedia)!: development of a structured wiki-based encyclopedia within a local austrian context. In *Int. Sym. Wikis* (2010).