

Quantifying the Accuracy of Relational Statements in Wikipedia: A Methodology

Gabriel Weaver
The Perseus Project
124 Eaton Hall
Medford, MA
gweave01@tufts.edu

Barbara Strickland
The Perseus Project
124 Eaton Hall
Medford, MA
bstric01@tufts.edu

Gregory Crane
The Perseus Project
124 Eaton Hall
Medford, MA
gcrane@tufts.edu

ABSTRACT

The Perseus Project at Tufts University produces tools to enhance the study of humanities texts. Perseus' new named-entity browser lets users browse an index of references to people, places, and dates within a work. A more advanced version of this same tool would enable users to navigate a work by specifying relations between two entities. Such tools require sets of relational data to function. This paper discusses a methodology for evaluating data sets obtained from Wikipedia and its potential.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval;

H.5.4 [Information Systems Applications]: Hypertext/HypermediaArchitectures

General Terms

design, experimentation

Keywords

Wikipedia, named-entity recognition, link analysis

1. INTRODUCTION

A great deal of recent research has focused on the quality of the information contained within Wikipedia [3]. In fact, a recent comparison of Wikipedia to *Encyclopedia Britannica* by Nature found that among 42 scientific entries compared by experts the difference in accuracy was not particularly great [2]. Less research has focused on the link structure of Wikipedia, particularly on the accuracy of the disambiguating links, although recent work by [1] proposes a methodology for discovering potential missing relevant links between Wikipedia articles. The combination of Wikipedia's large user base, the variety of topics it covers, and the ability to link from

one article to another make it a good source for relational data sets.

Wikipedia's internal links provide a data structure for disambiguating named-entities. An internal link associates a group of words with another Wikipedia article. For example, Wikipedia's article on Tufts University states "Charles Tufts founded Tufts College." In this relational statement, the subject is Charles Tufts, the object is Tufts College, and "founded" relates these two entities. The words "Charles Tufts" link to an article about Charles Tufts (1781-1876). An internal link can uniquely identify a person because a title in Wikipedia can only refer to one article. Two exceptions to this observation occur when the linked article is either empty or a disambiguation page. When an article is empty, it may not be clear what the article identifies without content. When an article is a disambiguation page, then the unique title references many different topics and so fails to uniquely identify one topic. An example of this would be if the "Charles Tufts" internal link linked to a page that listed many different Charles Tufts. An internal link to the title of a disambiguation page fails to uniquely identify the words associated with the link.

Disambiguating internal links can uniquely identify the subject and object of a relational statement within Wikipedia. In the following section, a methodology for studying the accuracy of these relations will be proposed. Finally, results from a preliminary investigation of Wikipedia and its accuracy for relations whose subjects are people, places, and organizations will be presented and discussed.

2. METHODOLOGY

A relational statement within Wikipedia is only accurate if the statement made is true and the articles referenced by the subject and object are correct. False statements are not suitable for use in a data set. Similarly, if the subject or object link to the wrong articles, then the statement is inaccurate. This informal definition of accuracy forms the basis of the proposed methodology used to evaluate Wikipedia.

For our preliminary study, we decided to examine the accuracy of relations involving people, places,

and organizations. The methodology separates data collection and interpretation. For the data collection step, we (1) signed into Wikipedia with a user account and (2) used the ‘Random Page’ feature to sample articles on various people, places, and organizations. For each article we (3) found a maximum of three statements about the subject of the article and recorded the following information: the relation, the display text, the article linked to, the date, the accuracy of the link for the object of the relation, and the accuracy of the relational statement itself. The accuracy of both the internal link and the relational statement were recorded with respect to the situations that could lead to inconsistencies in recording the accuracy of relational statements. External research and web resources were used to verify the accuracy of the relational statements. In total, we collected data for 200 relational statements: 119 on people, 49 on places, and 32 on organizations. These statements came from 73 different articles. Finally, in an optional step we (4) added each article reviewed to our account’s “watchlist.” This will allow us to study changes in the Wikipedia articles sampled over time and give us an idea of the continuing accuracy of a data set generated from Wikipedia without repeating the procedure.

3. RESULTS AND CONCLUSION

The accuracy of the relational data in Wikipedia depended upon the truth of both the relational statement and the linked article that denoted the object of this statement. These measurements are captured in Tables 1 and 2. For those statements that we could validate with an external source, only 5 out of 180 relational statements were false. The truth of some statements could not be validated. There was a total of 20 unvalidated statements for people, places, and organizations. An example of this occurred in an article on Osorkon III, who supposedly “defeated opposing forces of Shosheng VI.” Depending upon how important random sampling is to your evaluation, it may be helpful to select articles whose subjects are easily researched. This may bias the evaluation however, since such easily researched subjects are more likely to have accurate information available.

Entity	True	False
People	102 (98.08%)	2 (1.92%)
Places	43 (95.56%)	2 (4.44%)
Orgs	30 (96.77%)	1 (3.23%)
Total	175 (97.22%)	5 (2.78%)

Table 1: Measurements of the truth of relational statements within Wikipedia.

Table 2 shows the correctness of internal links within the relational statements studied. Only one article was incorrectly linked to in 200 statements. Furthermore, there were only 10 statements for which the correctness of the disambiguating internal link was undefined. The correctness of an internal link to a disambiguation page is undefined because the object of the statement is not uniquely identified. Such

links are not suitable for disambiguation. Similarly, the correctness of an internal link to an empty article is also undefined because, although an object is uniquely referenced by the article’s title, the object itself is undefined without any content. Depending upon how the data set is going to be used, these situations may or may not affect the accuracy of your evaluation.

Entity	Correct	Incorrect
People	109 (99.09%)	1 (0.91%)
Places	48 (100.00%)	0 (0.00%)
Orgs	32 (100.00%)	0 (0.00%)
Total	189 (99.47%)	1 (0.53%)

Table 2: Measurements of the correctness of disambiguating internal links in Wikipedia.

The accuracy of relational statements involving certain entity-types determines whether or not Wikipedia can be used as a data source for any tools seeking to leverage such information. The results gleaned from our 200 statements hint that Wikipedia might provide data suitable for constructing a set of relational data of people, places, and organizations. If one was going to sample more than 200 statements, it might be a good idea to record more than three relational statements per article, for less switching between articles would be involved. While this paper proposes a basic methodology for quantifying accuracy, it does not consider other variables which may be relevant to measuring the accuracy of a data set based upon Wikipedia. How does page activity or the number of people collaborating on an article affect the quality of statements? Are statements containing links with ambiguous display text more likely to be incorrect? In other words is the statement, “Lee went to Washington” more likely to be incorrect than “George Lee went to Washington state”? While these features of relational statements were not taken into account, the methodology proposed is a solid foundation upon which to base further study of whether or not wikis can provide data sets of relational information to help users navigate humanities texts.

4. ADDITIONAL AUTHORS

Additional authors: Alison Jones (The Perseus Project, email: ajones06@perseus.tufts.edu)

5. REFERENCES

- [1] S. Fissaha Adafre and M. de Rijke. Discovering missing links in wikipedia. In *LinkKDD-2005*, 2005.
- [2] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–901.
- [3] F. B. Viegas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582, 2004.