Computing Semantic Relatedness from Human Navigational Paths on Wikipedia

Philipp Singer
KTI, University of Technology
Graz, Austria
philipp.singer@tugraz.at

Thomas Niebler
DMIR, University of Würzburg
Würzburg, Germany
thomas.niebler@uni-wuerzburg.de

Markus Strohmaier
KTI, University of Technology
Graz, Austria
markus.strohmaier@tugraz.ac.at

Andreas Hotho
DMIR, University of Würzburg
Würzburg, Germany
hotho@informatik.uni-wuerzburg.de

ABSTRACT

This paper presents a novel approach for computing semantic relatedness between concepts on Wikipedia by using *human navigational paths* for this task. Our results suggest that human navigational paths provide a viable source for calculating semantic relatedness between concepts on Wikipedia. We also show that we can improve accuracy by intelligent selection of path corpora based on path characteristics indicating that not all paths are equally useful. Our work makes an argument for expanding the existing arsenal of data sources for calculating semantic relatedness and to consider the utility of human navigational paths for this task.

Categories and Subject Descriptors

H.1 [**Models and Principles**]: Miscellaneous; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia—*Navigation*

Keywords

semantic relatedness; navigation; Wikipedia

1. INTRODUCTION

Computing semantic relatedness between concepts represents a fundamental challenge on our way to a semantically-enabled web. While existing semantic analysis methods have shown great potential by using textual or structural (link) information on Wikipedia, they only capture semantics from a limited set of people (e.g., Wikipedia editors) and they mostly neglect pragmatics (i.e., how Wikipedia is used). At the same time, millions of web users navigate Wikipedia daily to find information, to educate themselves or for research issues. When navigating a set of articles on Wikipedia, users typically need to tap into their intuitions about real-world concepts and the perceived relationships between them in order to progress towards their set of targeted articles. West et al. [2] have been the first to study semantics in human navigational paths. While their work demonstrates the great potential of this approach, it is limited in some ways: (1) the distance between two concepts can only be computed if they directly co-occur in a path, (2) the work is limited to a small set of concepts and paths and (3) evaluation was performed by a few human judges only. In our work we cover all these problems and furthermore also provide insights into the usefulness of specific paths for the task of semantic relatedness calculation.

Consequently, we not only want to investigate (i) the usefulness of human navigational tasks for calculating semantic relatedness, but also (ii) which kinds of navigational paths are particularly useful? To tackle these questions we present a series of principled experiments studying the usefulness of almost 1.8 million human navigational paths on Wikipedia obtained from "TheWikiGame"[1] for determining semantic relatedness between concepts. In this game the goal is to navigate from one given *start page* to a given *target page*. For calculating characteristics of paths we also obtained the underlying topological link network of Wikipedia.

2. SEMANTICS OF PATHS

For determining semantic relatedness between concepts, we use *second-order co-occurrence* information within paths and calculate the *cosine similarity* between co-occurrence vectors, given two Wikipedia concepts. To capture relatedness of two concepts in a corpus of human navigation paths, we use *sliding windows* of a variable size k over the path sequence. Thereby, we follow the natural assumption that the distance between two concepts in navigational paths is crucial for calculating precise semantic relatedness scores.

Figure 1 illustrates how we calculate the co-occurrence between concepts available in a path with a sample window size of $k = 3$. Circles represent Wikipedia articles, rounded rectangles represent a window, the solid arrows represent the path taken and the dashed lines with dotted ends each represent a (symmetric) co-occurrence between two concepts. The final co-occurrence information is then projected to a large matrix where each row represents the co-occurrence vector of a specific concept, which is then used for the cosine similarity calculation.

To evaluate semantic relatedness, we compare our results to a golden standard dataset, specifically the *WordSimilarity-353* dataset [1]. It consists of 353 pairs of English words and names which have been judged by humans according to their semantic relatedness. We have manually mapped the word pairs to Wikipedia concept pairs.

[1] http://www.thewikigame.com

Copyright is held by the author/owner(s).
WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

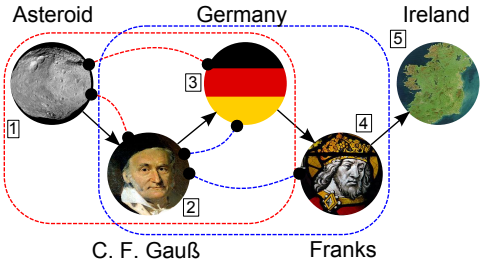

Figure 1: Demonstration of sliding windows

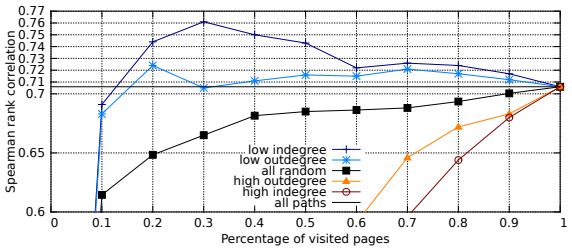

Figure 2: Semantic relatedness of selected path subsets

For evaluation we first calculate the semantic relatedness scores for each pair using our method and dataset and finally calculate the accuracy using the *Spearman rank correlation*.

Table 1 depicts the achieved results using the set of ≈1.8 million human navigational paths and indicates that the best accuracy can be achieved using a window size of 3 or 4[2]. The results demonstrate that human navigational paths contain information relevant for calculating semantic relatedness between concepts by exhibiting high quality relatedness evaluated against WordSimilarity-353.

Table 1: Semantic relatedness accuracy

Window size	None	2	3	4	5
Accuracy	0.644	0.636	0.706	0.713	0.686

3. PATH SELECTION EXPERIMENTS

Intuitively, paths containing more specific concepts can produce more precise and fine-grained semantic relatedness scores using the proposed co-occurrence information in comparison to more general paths. Such specific concepts should exhibit a lower in-degree value calculated on the underlying topological link network than concepts with more abstract patterns. We now aim to extract smaller sets of paths based on this idea and investigate whether these paths can outperform the accuracy using the set of all paths. To gauge the average abstractness of any path p in our corpus of paths $\mathbb{P}$, we measure the mean **in- and outdegree** of all nodes in a path and rank all paths based on their characteristics in both ascending and descending order. For each characteristic, we calculate ten subsets of increasing size where the tenth subset corresponds to the set of all available Wikigame paths. The sizes of our subsets are calculated by the number of visited nodes inside the subset. Thus, a potential path subset with very long paths consists of fewer actual paths than a set with mostly short paths, but both contain roughly the *same amount of visited nodes*. We also create a *random baseline* for each individual split by shuffling the original set of all paths and extracting subsets according to the selection process described above. After the generation of the path ordering lists and path selection, we run our semantic evaluation for each of these subsets.

In Figure 2, we present the semantic relatedness accuracy obtained from individual sub-corpora of navigational Wikigame paths using our selection strategies. The horizontal black line with a Spearman rank correlation of 0.706 shows the results achieved when taking a corpus of all Wikigame paths. A first observation is that we indeed can select smaller corpora of navigational paths that perform equally or better than the complete corpus of Wikigame paths. By incrementally adding paths with the *lowest average in-degree* of their concepts, we can achieve the highest Spearman rank correlation with a sub-corpus of only 30% of all Wikigame paths (dark blue line, plus symbol). The respective accuracy of 0.761 outperforms the accuracy of the whole Wikigame corpus by about 6% while covering less than a third of all visited pages in the complete corpus. Similar observations can be seen by filtering according to the *lowest average out-degree paths* (blue line, star symbol). Opposite filtered path corpora perform worse than the random baseline and the accuracy of all paths.

[2]For tractability and simplicity we focus on a window size of 3 for the rest of this paper.

4. DISCUSSION AND CONCLUSIONS

In this work we systematically evaluated information on *≈1.8 million human navigation paths* for calculating semantic relatedness between Wikipedia concepts. Our experiments show further indicators that (1) human navigational paths may provide a viable source for calculating semantic relatedness between concepts in information networks and (2) we find that not all navigational paths are equally useful. Intelligent selection of navigational paths based on path characteristics can improve accuracy. A limitation of our work is that the paths at hand are produced by a game and hence, may bias our obtained results. Nevertheless, they represent an abstraction of real user navigation in information networks, which we also want to investigate in future work. Our observed results suggest that in addition to data from textual or structural (link) sources, *usage* data - such as human navigational paths - could play a pivotal role in the future. Hence, we can envision that future methods for computing semantic relatedness might not produce objective scores for semantic relatedness, but *subjective* scores that take into account how concepts are used and perceived by large user populations.

5. ACKNOWLEDGEMENTS

We want to thank Alex Clemesha (TheWikiGame) for access to the TheWikiGame dataset and Denis Helic (TU Graz) for his cooperation. This work is in part funded by the FWF Austrian Science Fund Grant I677 and the Know-Center Graz as well as by the DFG through the PoSTS project.

6. REFERENCES

[1] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1):116–131, Jan. 2002.

[2] R. West, J. Pineau, and D. Precup. Wikispeedia: An Online Game for Inferring Semantic Distances between Concepts. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence*, IJCAI '09, pages 1598–1603, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.