WikiBrain: Making Computer Programs Smarter with Knowledge from Wikipedia

Toby Jia-Jun Li lixx2211@umn.edu **UROP Mentor: Brent Hecht** bhecht@cs.umn.edu
Department of Computer Science & Engineering, University of Minnesota Twin Cities

UNIVERSITY OF MINNESOTA
Driven to Discover℠

grouplens

Introduction

Abstract
Containing over 32 million pages in 287 languages, Wikipedia has become the largest ever repository of encyclopedic knowledge. In this work, we present *WikiBrain*, a software library we built that allows researchers and practitioners to incorporate intelligence from Wikipedia into their projects. *WikiBrain* not only allows users to access basic knowledge structures like the link graph and article text, but also incorporates powerful algorithms from the field of artificial intelligence that allow computers to understand **complex semantic relationships, spatiotemporal dynamics and cross-lingual linkages.**

As a proof of concept, we are employing *WikiBrain* to better understand the First Law of Geography, which states that, “everything is related to everything else, but near things are more related than distant things.” We are investigating the specific character of the association between distance and relatedness and are studying the roles of borders, time and culture in this association.

Semantic Relatedness Analysis
Semantic relatedness is a metric that gives a numerical relatedness score for any pair of lexically expressed concepts. It is widely used in the field of **artificial intelligence, natural language processing and computational linguistics**. It is also a good heuristic for other kinds of relatedness.

WikiBrain provides a set of easy-to-use SR tools, including implementations for the *ensemble* metric, the *inlink* metric and modules to find the most similar concepts etc.

Figure 1: The visualization of seveal most semantically related concepts to “Cat” in Simple Wikipedia using the *inlink* metric

Highlight Features

Spatial Analysis
WikiBrain provides spatial tagging tools and references for Wikipedia entities. Users can easily connect the concepts in Wikipedia to not only the geographical reference system, but **all kinds of spatial reference systems.**

By connecting Wikipedia concepts to geometries in spatial reference systems, users can apply a huge set of spatial analysis operations on the Wikipedia data like **calculating spatial distances, differences, intersections, unions and convex hull of Wikipedia-tagged geometries.**

Spatial Analysis also provides **great visualizations** for Wikipedia knowledge. Below is an example of visualizing the semantic relatedness between Wikipedia entities on different spatial reference systems.

Figure 2: The visualization of the sematic relatedness between the concept “Nuclear Power” and the elements on the “Periodic Table” Reference [1]

Figure 3: The visualization of the sematic relatedness between the concept “Nuclear Power” and the countries on the “World Map” Reference [1]

Multi-Lingual Support
Wikipedia is a source of world knowledge with great cultural diversity across different language editions. Previous works have shown that **over 74% of concepts** are described in only one language and **only around 0.12% concepts** are represented in all 25 major language editions of Wikipedia [2]. Even for these “globally relevant” concepts, the topics and contents covered in each entry of different language **varies a lot** [2] [3].

WikiBrain references articles **hyper-lingually** by nature, which supports **cross-language analysis** of Wikipedia entries. Users can **simultaneously** access the entries of the same concept from different language editions.

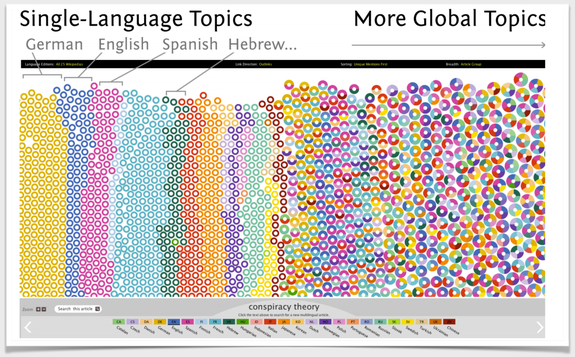

Figure 4: A visualization of topics covered in the multi-lingual article “Conspiracy Theory” [3]

Wikidata Support
WikiBrain supports parsing and analyzing data from Wikidata, which is a project started by Wikimedia to create a structured machine-readable database of knowledge.

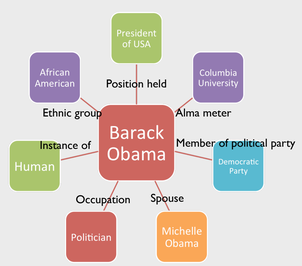

Figure 5: A simplified visualization of Wikidata item “Barack Obama”

With the Wikidata support, *WikiBrain* is now able to recognize the **type of relationships or connections** between Wikipedia concepts.

Research with WikiBrain

First Law of Geography

“Everything is related to everything else, but near things are more related than distant things”
-Waldo R. Tobler

We are evaluating the First Law of Geography [4] in the domain of Wikipedia knowledge using the *WikiBrain* library. Through examining the spatial distance between spatial-tagged multi-lingual Wikipedia concepts and comparing the spatial distances with the corresponding semantic relatedness. Previous work [5] has shown that the First Law of Geography is generally true in the domain of massive, domain-neutral representation of world knowledge. We are interested in using *WikiBrain* to find the effects of other factors like **temporal distance and cultural, geographical or political borders** in addition to the spatial distance on the semantic relatedness between concepts.

References

[1] Hecht, B., Carton, S., Quaderi, M., Schöning, J., Raubal, M., Gergle, D., Downey, D. Explanatory Semantic Relatedness and Explicit Spatialization for Exploratory Search. Proceedings of ACM SIGIR 2012. New York: ACM Press.

[2] Hecht, B. and Gergle, D. The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context. Proceedings of CHI 2010, pp. 291–300. New York: ACM Press.

[3] Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M. and Gergle, D.. Omnipedia: Bridging the Wikipedia Language Gap. Proceedings of CHI 2012. New York: ACM Press

[4] Tobler, W.R.: A computer movie simulating urban growth in the Detroit region. Economic Geography 46 (1970) 234-240

[5] Hecht, B. and Moxley, E. Terabytes of Tobler: Evaluating the First Law of Geography in a Massive, Domain-Neutral Representation of World Knowledge. Proceedings of the 2009 International Conference on Spatial Information Theory , pp. 88-105. Berlin: Springer-Verlag.

This project was supported by the University of Minnesota's Undergraduate Research Opportunities Program