From GPS and Google Maps to Spatial Computing

VGI Module Programming Assignment

Assignment by Toby J. Li and Dr. Brent Hecht

As discussed in the videos, one of the most influential types of VGI is Wikipedia VGI, which consists of geotagged Wikipedia articles and their connections to other articles and resources in Wikipedia. In this programming assignment, we will use WikiBrain[5] (www.wikibrainapi.org) to access the incredible and powerful new geographic dataset that is Wikipedia VGI.

Important: All answers should be submitted online in Coursera. See the Module 5 page for instructions.

Prior to completing the tasks below, you must do the following:

- Watch the video on the programming assignment (the last video in the module).
- Set up the WikiBrain environment according to the installation instructions.
- Run and explore the example program provided at org.wikibrain.cookbook.mooc.QueryExample (located at wikibrain/wikibrain-cookbook/src/main /java/org/wikibrain/cookbook/mooc/QueryExample.java)
- We highly recommend that you read the paper "WikiBrain: Democratizing computing on Wikipedia" by Sen et al. [5] before starting the assignment
- We also highly recommend you review the documentation on WikiBrain's website (http://www.wikibrainapi.org).

Important: The QueryExample.java file noted above will GREATLY assist you in completing the tasks below. Please be sure to review this file closely before starting and refer back to it as you go through the assignment.

Important: Be sure to enter all answers that are strings *verbatim* from WikiBrain's output. We will not be able to accept variations on these strings as correct answers. So, for example, if the answer is "Minnesota", do NOT enter anything like "MN" or "Minn."

All of these tasks use the Simple English Wikipedia, which is a language edition of Wikipedia intended for English language learners. It contains around 115,000 articles, as opposed to the "full" English Wikipedia's millions of articles. As such, Simple English is commonly used in Wikipedia-based research and practice as a way to develop and test applications on a small dataset prior to the use of full English.

1. Basic spatial query:

Note: For this question and for the rest of the assignment, most of the code for these tasks is already written in the QueryExample.java file. You just need to find the right snippet of code and adapt it for the needs of each task.

a. How many geotagged Simple English articles are there located China that have a title that includes the word "University" or "College"?
b. Among the above universities and colleges, which one has the shortest straight-line distance to the University of Minnesota? (Hint: The University of Minnesota has a geotagged Wikipedia page). Please enter only the article title, not the language edition, etc.

2. Semantic Relatedness (SR) is a metric that assesses the number and strength of relationships between two entities. Wikipedia is now widely used as a data source for computing SR [1,3,6], and we have built a few state-of-the-art SR algorithms into WikiBrain. WikiBrain's SR algorithms produce a value between 0 and 1 for any two input Wikipedia articles (describing entities), where a larger number indicates stronger relatedness. We'll cover SR in more detail in module 6, but the power of SR algorithms and their application to VGI should be clear in the following two tasks.

Note: Be sure to use the correct case in the titles below.

a. According to the ensemble SR metric running on Simple English Wikipedia, which of the following article pairs is the most related?
    A. "Cat" and "Dog"
    B. "Hamburger" and "French fries"
    C. "New York City" and "San Francisco"
    D. "Cat" and "Elephant"

b. What is the most related geotagged Simple English article to "Pizza" in Germany? (Please only enter the article name, not the language edition, etc.)

3. The **indegree sum** (i.e. the sum total of the links pointed to a set of articles) for all geotagged articles in an area is a powerful metric. It is both an indicator of how much that area is discussed in Wikipedia and, as a *centrality measure*, it is a proxy of how *important* that region is according to Wikipedia [4].

We're going to calculate the indegree sum in the Simple English Wikipedia for several states in the United States. Please follow the steps below, and then answer questions (a) and (b).

i. For each of the following U.S. states, get a list of all geotagged Simple English articles located in that state from the layer "wikidata": California, Minnesota, Illinois, Florida, West Virginia.
ii. Count the inlinks to all geotagged articles in each state. The total number of inlinks pointing to geotagged articles in each state is the state's **indegree sum**. (Note: for inlinks, get parseable inlinks exactly as is depicted in Part 6 of the QueryExample.java file. No need to modify the getLinks() call other than to change the article).

a. Which of the five states has the highest indegree sum?
b. Which of the five states has the lowest indegree sum?
c. How many inlinks does the "University of Minnesota" article contribute to Minnesota's indegree sum?

Please see Hecht and Gergle [4] and Hecht [2] for more information about indegree sums and their utility.

Acknowledgements

We would like to thank Dr. Shilad Sen , Isaac Johnson, and Yilun (Allen) Lin for their help developing this assignment.

References

1. Egozi, O., Markovitch, S., and Gabrilovich, E. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS) 29*, 2 (2011), 8.
2. Hecht, B. The Mining and Application of Diverse Cultural Perspectives in User-Generated Content. 2013.
3. Hecht, B., Carton, S.H., Quaderi, M., et al. Explanatory semantic relatedness and explicit spatialization for exploratory search. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, ACM (2012), 415–424.
4. Hecht, B. and Gergle, D. Measuring self-focus bias in community-maintained knowledge repositories. *Proceedings of the fourth international conference on Communities and technologies*, ACM (2009), 11–20.
5. Sen, S., Li, T.J.-J., WikiBrain Team, and Hecht, B. WikiBrain: Democratizing computation on Wikipedia. *Proceedings of the 10th International Symposium on Open Collaboration (WikiSym + OpenSym 2014)*, ACM.
6. Witten, I. and Milne, D. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, (2008), 25–30.