Brent Hecht
VGI Position Statement

Wikipedia Volunteered Geographic Information

A common definition of geographic information is <**x**,**z**>, where **x** is some location in space-time and **z** is some set of general properties, or attributes. My interests in volunteered geographic information (VGI) lie much more in the **z** than the **x**. Broadly stated, I explore the character and applications of the *attributes* of massive repositories of VGI. Although I am beginning to investigate VGI attributes from social networks such as Facebook, most of my work thus far has been researching the uniqueness and application potential of the attributes of Wikipedia VGI.

All Wikipedia spatial articles have massive amounts of attribute information associated with them. This attribute data is comes in a variety of structural forms, from natural text to nodes and edges within a graph structure. The one commonality these structures have is that they are atypical for GIScience use. I have identified three attribute structures that have proven to be very interesting phenomenon and fruitful for novel applications: the Wikipedia Category Graph (WCG), the Wikipedia Article Graph (WAG), and the Wikipedia Natural Text (WNT). I have used one or all of these structures in several research projects, the papers on which are listed at the end of this statement. Rather than describing the research, however, I will discuss the structures themselves so as to help stimulate discussion with, and possibly future work by, my fellow VGI researchers.

Wikipedia VGI is unique mainly *because* of the attribute information that each spatial coordinate contains. However, in order to understand the value of Wikipedia VGI attributes, it is first important to understand the context under which spatial coordinates are inputted to the encyclopedia. Wikipedia articles are spatially referenced by Wikipedia users through a collaborative geotagging process. On an implementation level, this process is executed in Wikipedia entirely through the use of templates. Templates are delimited with opening and closing double curly-braces (i.e. “{{template}}”) in WikiScript and essentially describe a function name and its parameters. Wikipedia spatial reference templates can be used in two very different ways, which can sometimes be differentiated by the template chosen and sometimes by the parameters of the template. The mostly widely employed usage is to provide a solitary spatial reference, with the semantic value of the reference applying to the entire article. It is these articles that I have termed *spatial articles*. However, the same templates (with different parameters) or very slightly modified templates are also used quite often within the body of an article to describe a spatial location inline. Inline templates do not represent spatial articles as we have defined them here, as they do not reference the entire article but rather the text in which they are embedded.

Since version 1.3 of the MediaWiki software was released in May 2004, each Wikipedia has had its own WCG (Voss 2006). In many of the major Wikipedias (the standard terminology is to refer to each language version of Wikipedia as a different Wikipedia), the vast majority of the articles are nodes in the WCG. To establish an article as a node in this graph, a Wikipedian must simply tag the article with category information. In the English Wikipedia, this means adding a link to the article in the format of [[Category:CategoryName]]. Other Wikipedias have very similar syntax, replacing the word “Category” for its translation to the Wikipedia’s native language. Each article can have none, one, or many category memberships. Clicking on these category links forwards users to category pages, which themselves can be tagged with category information, making them into sub-categories. This hierarchical tagging regime has resulted in a pseudo-taxonomy of categories which can get quite large. The October 2007 German WCG had a total of 45,636 vertices and 82,584 edges. Voss (2006) and Strube and Ponzetto (2006) identify the WCG as a “folksonomy”. VanderWal (2004), who is credited with the term “folksonomy”, defined his label as the “bottom-up social classification that takes place on Flickr, del.icio.us, etc.” Unlike Flickr and del.icio.us, the WCG folksonomies can be hierarchical (as noted above) and, as such, have been defined as thesauri (Voss 2006). The WCGs are also unique in that all tags must be implicitly agreed upon by all users in the community; the tagging strategy is thus a collaborative one. According to Voss (2006), the WCGs represent the first-ever information store that includes both thesauri and collaborative tagging.

The WAG can be defined as WAG = (*A*,*L*), where *A* is the set of articles in a given Wikipedia and *L* is the set of standard links between these articles. Formally, graphs are usually defined as an ordered triple, where a graph *G* = (*V*, *E*, □) . *V* is the set of vertices in the graph, *E* is the set of edges, and □ is the “edgemap” that defines which members of V form the endpoints of each edge in *E* (Agnarsson and Greenlaw 2007). In Wikipedia, *A* = *V* and *L* = *E*. The endpoints of each edge in *E* is implicit to the definition of each edge, which must be defined by Wikipedians as a link from one article to another. As such, there is no explicit □ structure. While, the size of *A*, or |*A*|, and the size of *L*, or |*L*|, varies greatly from Wikipedia to Wikipedia, for the larger Wikipedias, the WAG is enormous. In the latest Wikipedia data dumps used for my research projects, which were generated in October 2007, the English Wikipedia had |*A*| ~= 2.05 million and |*L*| ~= 45 million and the German Wikipedia had had |*A*| ~= 0.69 million and |*L*| ~= 15.0 million. The size of the graph creates certain challenges and forces long processing times, issues that are important to consider when doing WAG-based research. Another key feature of the WAG is that its links are replete with non-classic relations (Morris and Hirst 2004). The immense utility of this characteristic are discussed in my research papers listed at the end of this statement.

The Wikipedia Text (WT) data source is defined as all natural text that occurs on the article pages, with the exception of text that occurs in link targets with alternative labels and text that occurs in templates. The snippet sub-structure is probably the most important substructure of the WT, at least in the context of my research. First identified in (Hecht et. al 2007), a snippet is a paragraph in the WT between *n* and *m* characters (*n* and *m* are set based on the needs of a particular task) that is by delimited one or more new line characters. Text that is a member of any titles is excluded. The Wikipedia snippet is a unique natural text phenomenon in that we have found qualitatively that nearly all snippets are entirely independent of other snippets within the same article. In other words, snippets rarely contain ambiguous text that the reader is expected to disambiguate using knowledge acquired from other snippets on the same Wikipedia page. This is important because snippets can be safely rearranged or presented on their own without severely reducing their information content. This property of snippets is used in every Wikipedia research project in which I have participated. While my work mainly uses subsets of the WT, the WT in its near entirely is used by some researchers, mostly as a source for a distributional natural language processing methodologies. In other words, the WT resource makes excellent bag-of-words vectors that can be used to describe the subject of Wikipedia articles.

My future research will involve further exploring properties and applications of the above attribute structures, as well as the investigation of other structures, such as the multi-lingual extension of the WT. I also am beginning to research the copious natural text and graph structure attributes of Facebook VGI, for which the spatial information lies in users' submission of their current location, hometown location, current place of work or educational institution, etc.

Further Reading

Hecht, B. Using Wikipedia as a Spatiotemporal Knowledge Repository. Geography Masters of Arts (2007).

Hecht, B. & Raubal, M. GeoSR: Geographically explore semantic relations in world knowledge. AGILE 2008 (2008). *IN SUBMISSION.*

Hecht, B., Rohs, M., Schöning, J. & Krüger, A. WikEye - Using Magic Lenses to Explore Georeferenced Wikipedia Content. PERMID 2007 (in conjunction with the Fifth International Conference on Pervasive Computing) 6-10 (2007).

Hecht, B., Starosielski, N. & Dara-Abrams, D. Generating Educational Tourism Narratives from Wikipedia. Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium on Intelligent Narrative Technologies 37-44 (2007).

Schöning, J., Hecht, B., et al. Improving Interaction with Virtual Globes through Spatial Thinking: Helping Users Ask "Why?". Intelligent User Interfaces 2008 (IUI 2008). *IN PRESS*.

Schöning, J., Hecht, B., Rohs, M. & Starosielski, N. WikEar − Automatically Generated Location-Based Audio Stories between Public City Maps. 9th International Conference on Ubiquitous Computing Demo Proceedings (2007).

Bibliography

Agnarsson, G. & Greenlaw, R. Graph Theory: Modeling, Applications, and Algorithms (Prentice Hall, 2006).

Hecht, B., Starosielski, N. & Dara-Abrams, D. Generating Educational Tourism Narratives from Wikipedia. Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium on Intelligent Narrative Technologies 37-44 (2007).

Morris, J. & Hirst, G. Non-classical lexical semantic relations. Workshop on Computational Lexical Semantics, Human Language Technology Conference of the North American Chapter of the Assocation for Computational Linguistics (HLT-NAACL 2004) (2004).

Voss, J. Workshop on Wikipedia Research. WikiSym 2006 127 (2006).

Strube, M. & Ponzetto, S. P. WikiRelate! Computing Semantic Relatedness Using Wikipedia. AAAI 2006 (2006).

VanderWal, T. You Down with Folksonomy. (2004).