GeoSR: Geographically Explore Semantic Relations in World Knowledge

Brent Hecht and Martin Raubal

Department of Geography
University of California, Santa Barbara
1832 Ellison Hall, Santa Barbara, CA 93106-4060
{bhecht, raubal}@geog.ucsb.edu

Abstract. Methods to determine the semantic relatedness (SR) value between two lexically expressed entities abound in the field of natural language processing (NLP). The goal of such efforts is to identify a single measure that summarizes the number and strength of the relationships between the two entities. In this paper, we present *GeoSR*, the first adaptation of SR methods to the context of geographic data exploration. By combining the first use of a knowledge repository structure that is replete with non-classical relations, a new means of explaining those relations to users, and the novel application of SR measures to a geographic reference system, GeoSR allows users to geographically navigate and investigate the world knowledge encoded in Wikipedia. There are numerous visualization and interaction paradigms possible with GeoSR; we present one implementation as a proof-of-concept and discuss others. Although, Wikipedia is used as the knowledge repository for our implementation, GeoSR will also work with any knowledge repository having a similar set of properties.

1 Introduction and Related Work

In today's information-overloaded world, researchers in both the academic and professional community, students, policy analysts and people in many other fields frequently find themselves in the position of trying to locate a useful needle of information in a haystack of data. This search is often aided by the use of a spatial lens, as up to 80 percent of human decisions affect space or are affected by spatial situations (Albaredes, 1992). For example, a student doing a project on Judaism, love, George W. Bush, Berlin or any other concept or named entity will definitely want to know the places that are most related to these concepts and named entities and why. *GeoSR* provides users with a novel method of easily accomplishing this task.

1.1 GeoSR and Wikipedia

GeoSR uses Wikipedia as its knowledge repository. The introduction of every paper produced by the burgeoning Wikipedia research community has its own way of describing the phenomenon that is Wikipedia. However, they all seem to agree on several vital properties. First, Wikipedia is a free encyclopedia that is produced via a collaborative effort by its contributors. Second, Wikipedia is highly multilingual, with hundreds of available languages. Third, Wikipedia is enormous and is, by far, the largest encyclopedia the world has ever seen. Indeed, as of October 2007, Wikipedias in 14 languages had over 100,000 articles and the largest Wikipedia, English, had over 2.05 million. Finally, many researchers argue that Wikipedia "has probably become the largest collection of freely available knowledge" (Zesch et al., 2007a, p. 1).

The above facts are all relatively well known among people who use Wikipedia, which in the U.S. amount to 36 percent of the Internet-using population (Rainie and Tancer, 2007). However, what is less understood in the general and scientific communities are the opportunities presented by the massive knowledge repository of ubiquitously available information that Wikipedia represents. The research here is part of the first work (Hecht, 2007) that explores the *spatio-temporal possibilities* of this knowledge repository, as well as others in the future that could offer similar content, structure, and size (for example, *Citizendium*[1]). Several authors have conducted other research projects in this area including Minotour (Hecht et al., 2007a), WikEye (Hecht et al., 2007b), and WikEar (Schöning et al., 2007a).

It is important to note that because this research uses Wikipedia as a data source, it is vulnerable to the risks of Wikipedia information as identified by Denning et al. (2005). However, we believe these risks apply only minimally to GeoSR for the following reasons: (1) GeoSR is not tied to the editorial policies of Wikipedia, only its structure and size and, as such, the research is much more general than the data set it relies on, (2) GeoSR provides a novel and useful method for visualizing and exploring data people are already accessing in massive numbers despite the risks, and (3) Giles (2005) has shown that the accuracy of Wikipedia, at least in the scientific context, is comparable to that of more conventional encyclopedias.

1.2 GeoSR and Semantic Relatedness

Semantic relatedness (SR), which is at the heart of GeoSR, is a well-known topic in the field of natural language processing (NLP). There are many applications of SR in NLP, including word sense disambiguation, text summarization, information extraction and retrieval, and correction of word errors (Budanitsky and Hirst, 2006). There are two general methodological families of SR measures; SR measures based on graph- or network-based lexical resources, from which this research derives inspiration, and SR measures based on distributional similarity, which implement bag-of-word techniques. However, it has been argued that the distributional similarity

1 http://www.citizendium.org

family "is not an adequate model for lexical semantic relatedness" (Budanitsky and Hirst, 2006, p. 30).

SR is often confused with semantic similarity. While many fields use the concept of semantic similarity differently, in the world of NLP, similarity measures are identical to SR measures if and only if the only relationships being examined are hypernymy and hyponymy (the *isA* relationship viewed from both sides). Similarity is thus a special case of SR (Budanitsky and Hirst, 2006).

While members of the NLP community have presented myriad SR measures, most of these are designed for WordNet (Miller, 1995), GermaNet (Kunze, 2004), or older knowledge repositories. Very recently, some researchers have been investigating the modification of these methods for Wikipedia. Wikipedia has three structures that can be used to measure semantic relatedness: the Wikipedia Category Graph (WCG), the Wikipedia Article Graph (WAG), and the text of the Wikipedia entries (WT) (see (Zesch et al., 2007a) and (Hecht, 2007)). Strube and Ponzetto (2006) presented the first effort to estimate SR using Wikipedia, *WikiRelate!*. It uses the WCG and reported slightly better correlation with human judgments – the so-called "gold standard" of SR measures, even though many researchers have taken issue with available datasets – than similar WordNet-based measures for some test sets.

Very recently, Gabrilovitch and Markovitch (2007) developed *Explicit Semantic Analysis* (*ESA*), which used the WT structure with much improved results over WikiRelate! (as well as methods developed using other knowledge repositories) in terms of correlation with the gold standard. However, ESA relies exclusively on distributional similarity mechanisms.

Both ESA and WikiRelate! use the English Wikipedia as its knowledge repository. Zesch et al. (2007b) compared GermaNet and the German WCG for use in semantic relatedness applications. They concluded that Wikipedia excels at SR, while GermaNet is better for similarity applications (as defined by the NLP community).

All of the aforementioned SR measures were designed for traditional NLP applications. Because of the data exploration needs of the GeoSR project and especially because of the importance of spatial-entity-to-spatial-entity and spatial-entity-to-non-spatial entity relationships, it was necessary to develop a novel SR measure and corresponding algorithm for this research. We have called this measure, which is the first to use the Wikipedia Article Graph (WAG), *ExploSR* (pn: "explosure").

1.3 Overview of Paper and System Framework

The framework of GeoSR is as follows: Wikipedia provides the world knowledge and the ExploSR semantic relatedness measure is responsible for assigning relative weights to the myriad relationships found in the Wikipedia repository. Based on some input named entity or concept (such as Judaism, love, George W. Bush, or Berlin), these values are then visualized *geographically* in one of several ways using *spatial articles* as anchors in a geographic reference system. Users can employ these visualizations as a context from which to engage in data exploration. Figure 1 demonstrates one possible visualization and interaction schema, which is discussed in

more detail in section five. Only the top 100 locations are shown. For the location "Tuxplan (Veracruz)" (in Mexico), the explanation information is found in the "Identify" window in the "Explanatio" field, and can be seen in greater detail in figure 2. This data has been generated using the German Wikipedia, with the "Explanatio" field manually populated with English information. Missing links have not been included in this iteration of GeoSR due to implementation issues that are discussed in section 4.2.

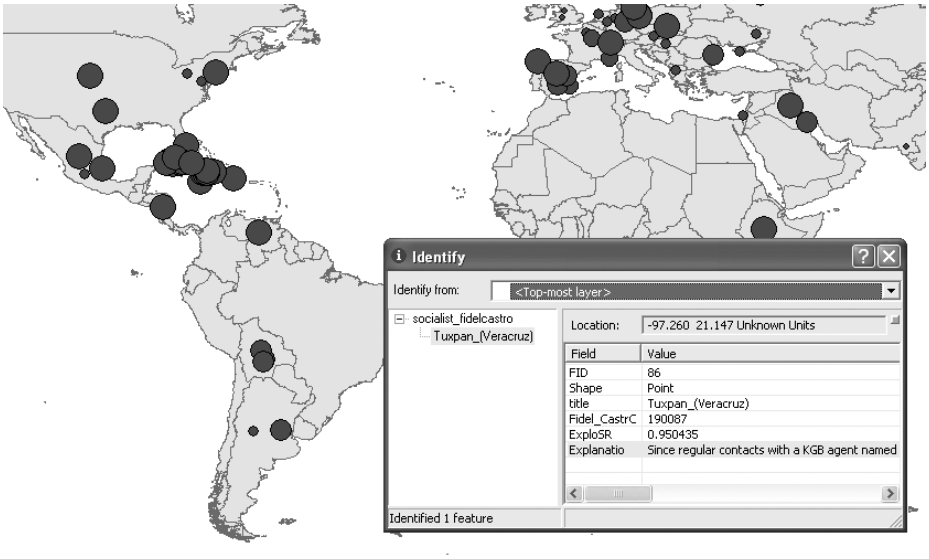

Figure 1: A visualization of GeoSR data in which *Fidel Castro* was the input entity. Large dots represent the most related locations to *Fidel Castro* and smaller dots represent less important locations (within the top 100 locations).

Fidel_Castro -> Tuxpan_(Veracruz)

Fidel Castro -> Cuban Revoluion -> 26th of July Movement

Since regular contacts with a KGB agent named Nikolai Sergeevich Leonov in Mexico City had not resulted in the hoped for weapon supply, they decided to go to the United States to gather personnel and funds from Cubans living there, including Carlos Prío Socarrás, the elected Cuban president deposed by Batista in 1952. Back in Mexico, the group trained under a Spanish Civil War Veteran, Cuban-born Alberto Bayo[35] who had fled to Mexico after Francisco Franco's victory in Spain. On November 26, 1956, Castro and his group of 81 followers, mostly Cuban exiles, set out from **Tuxpan, Mexico** aboard the yacht Granma for the purpose of starting a rebellion in Cuba.

Figure 2: An expansion of the content of the explanation field seen in figure 1.

Section two of this paper describes the pre-processing of Wikipedia required before its use in GeoSR and lays out a spatio-temporal framework with which to view

Wikipedia. The advantages of using the Wikipedia Article Graph (WAG) over other structures in the encyclopedia in this context are discussed in section three. Section four covers ExploSR in detail, highlighting its strengths and weaknesses. In section five, several applications for GeoSR are presented and one is fully demonstrated as a proof-of-concept. Finally, we wrap things up with a conclusion and describe directions for future research in section six.

2 Wikipedia Knowledge Repository

2.1 Preprocessing and API Access

While the Wikipedia knowledge repository has made implementing this research possible, a large number of pre-processing steps are necessary before Wikipedia can be used efficiently by GeoSR. Wikipedia data is received in the form of the XML "database backup dumps" provided by the Wikimedia Foundation[2], which runs Wikipedia. Dumps are made available every three to four weeks in every language in which there is Wikipedia. These dumps represent an enormous amount of text; the English Wikipedia dump from October 23, 2007 weighs in at 12.3 GB and the October 10, 2007 dump from second largest Wikipedia, that of German, is a sizeable 3.63 GB.

Once these dumps are downloaded, they must be processed by our Wikipedia parser and API, WikAPIdia, which we are considering releasing in the near future. During the parsing stage, structured information about each article including data about links, text, titles, title aliases (redirects), and much more is stored in a series of MySQL tables. Due to its size, the parsing step for the English Wikipedia can take a moderately-powered computer up to two to three days.

The MySQL database forms the data model from which the API portion of WikAPIdia operates. This API is the back end of all the Wikipedia-related projects in which our research group has participated, including GeoSR. It is important to note that while the only Wikipedias currently supported by our software are that of English, German and Spanish, we have constructed the software such that support for other Wikipedias is quite simple to add for a native speaker of that language.

2.2 Spatio-temporal Wikipedia data

In addition to processing lexical structures, WikAPIdia has special facilities for mining the spatial and temporal data in Wikipedia. Spatial data mainly comes in the form of explicitly "geotagged" articles, or articles with spatial reference information that describes the location of their subjects. We have labeled articles with spatial references as *spatial articles* and those without *non-spatial articles*. The distinction

2 http://www.wikimedia.org

between spatial and non-spatial articles plays a critical role in this research. Spatial articles are the intersection between "geographic space" and "Wikipedia space". As such, as will be explicated further in section five, spatial articles can essentially represent SR value samples in the real world.

The corollary to spatial articles in the temporal domain are what we call *pure temporal articles*, which, through their titles, contain references to a temporal reference system. While some of these references, such as the article titled "October 29" are ambiguous, others are not (such as "1983" or "April 29, 1983"). The pure temporal article construct plays an important role in our ExploSR algorithm (shown in section four), although its reference system utility is not emphasized in this research. Hecht (2007) provides a more general description of our Wikipedia spatio-temporal framework.

3 Advantages of the Wikipedia Article Graph

As noted in the introduction, ExploSR is the first SR methodology designed explicitly for data exploration use. However, it is also unique in that it is the first Wikipedia-focused measure to use the Wikipedia Article Graph (WAG). The WAG is the graph that is composed of the set of articles in a Wikipedia (set A), and the standard links between them (set L), which are defined using brackets in the Wiki markup language. Formally, graphs are usually defined as an ordered double, where a graph $G = (V, E)$. V is the set of vertices in the graph, and E is the set of edges (Piff, 1991). In this case, $A = V$ and $L = E$.

The WAG has two essential properties. First and foremost, the WAG is the ideal Wikipedia structure to use for data exploration SR measures because it is a simple matter to explicitly explain to users the relationships that resulted in the measure value between any two concepts. Secondly, the WAG contains much broader and deeper relation information than the knowledge repositories commonly used in SR research as well as other structures embedded in Wikipedia. This fact proves vital to examining relations between two spatial features and those between a spatial feature and non-spatial entity. The rest of this section is dedicated to explaining these two advantages in detail.

3.1 The Wikipedia Snippet – Paragraph Independence Facilitates Data Exploration

Nearly all articles in Wikipedia have uniquely independent paragraphs, which we term *snippets*. The Wikipedia snippet is a distinctive natural text phenomenon in that we have found qualitatively that nearly all Wikipedia 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 it signifies that the meaning of a link is almost always contained within the snippet that hosts the link (see figure 2). Additionally, this property ensures that snippets can

be safely rearranged or presented independently without severely reducing their understandable information content. We have found that the only context necessary for fully comprehending nearly all snippets is the title of the Wikipedia article in which they appear. Most of the remaining snippets can be completely framed by providing the hierarchy of titles, headings, and subheadings under which a snippet appears (i.e., for the *United States* article, *United States* -> History of the United States -> Revolutionary War).

Thus far, two possible causes of the unique snippet substructure in Wikipedia have been identified. The first is the collaborative nature of Wikipedia. Buriol et al. (2006) found that the average Wikipedia article has at least seven authors. This means that, in many cases, different parts of an article are written by different contributors, surely adding to the disjointedness of the text. This disjointedness, however, is desired in the Wikipedia community because of the encyclopedic nature of the writing style in Wikipedia. This writing style, termed *WikiLanguage* by Elia (2006), is the second identified cause of the independence of snippets. Wikipedians do not seek to create prose that flows from paragraph to paragraph; they seek to inform about facts in an organized fashion.

In summary, the independence of snippets provides an easy way to identify and present to the user the subset of text on any Wikipedia page that can explain the meaning of a link between two pages: the snippet in which the link resides. Explaining the meaning of links in the WCG in a similar manner would be impossible, as the meaning of WCG relationships is never explicitly explained. ESA, which is a distributional similarity measure, identifies relationships essentially by measuring the similarity between the unique words of Wikipedia articles. As such, using ESA to provide the full meaning of relationships between these articles in human-readable form would require a process entirely exogenous to the relatedness measure.

3.2 Depth and Breadth of Encoded Relations in the WAG

The second advantage of a WAG-based measure in the context of this research relates to the unique spatial needs of GeoSR. It has been qualitatively found that both ESA and WCG-based methods alone do not work well for spatial/spatial and spatial/non-spatial relationships. While this, along with the effectiveness of ExploSR outside the data exploration context, will be investigated in detail in future research, it is believed that the failure of WCG- and WT-based methods in the spatial context results from two characteristics of those two data structures: missing *classical relations*, and the worse offender, missing *non-classical relations*.

Morris and Hirst (2004) define classical relations as relations that depend on the sharing of properties of classical categories (Lakoff, 1987). Common classical relations include hypernymy/hyponymy (*isA*), meronymy/holonymy (*hasA*), synonymy (*likeA*) and antonymy (*isNotA*). WordNet, the lexical resource focus of most semantic relatedness research, offers only relations of this type. The vast majority of relations in the WCG are classical, and in fact are limited almost entirely to *isA* relations with a sprinkling of meronymy/holonymy (Zesch et al., 2007b). The WCG contains a large number of missing important *hasA* relations (not to mention

displaying a complete lack of antonymy, synonymy, etc.), making the WCG weak in both breadth and depth of classical relational coverage. In sum, the WCG is essentially a semantic similarity resource, not a SR resource (as defined by the NLP community). This is a critical problem when it comes to spatial entities: a hypernymy/hyponymy-only (*isA*-only) path in a taxonomy in which one endpoint entity is a spatial entity essentially limits the path to spatial entities. For instance, a spatial entity such as *California* is no doubt closely related to *Gold Rush*, but it is difficult to imagine a short hypernymy/hyponymy path between the two entities in a graph, even though the meronymy/holonymy relation is direct. Similarly, in the case of the WT, the unique word vectors of the *Gold Rush* article and that of the *California* article are highly dissimilar; the *Gold Rush* article focuses on the details of gold rushes in general and the California article is a broad overview of the state. As such, distributional measures also fail to understand the important *California-Gold Rush* meronymy/holonymy relation, which is captured at a simple path distance of 1 in the WAG.

Spatial/spatial and spatial/non-spatial article relationships also tend to display a large number of *non-classical* relations. Non-classical relations are associative or ad-hoc in nature (Budanitsky and Hirst, 2006) and are defined by Morris and Hirst (2004) as relations that "do not depend on the shared properties required of classical relations" (p. 2). Budanitsky and Hirst (2006) list the following examples of these types of relations: *isUsedTo* (*bed-sleep*), *worksIn* (*judge-court*), *livesIn* (*camel-desert*), and *isOnTheOutsideOf* (*corn-husk*). The WAG is absolutely replete with these types of relations. For instance, all of the above relations are encoded as at least unidirectional links in the English WAG (*judge-court* is bidirectional). Despite the fact that non-classical relations have been found to be an extremely important aspect of lexical relationships (Budanitsky and Hirst, 2006; Morris and Hirst, 2004), all graph-based SR research on Wikipedia thus far has focused on the WCG, which encodes almost none of these relations. The extent to which a distributional measure such as ESA understands non-classical relations is unclear.

Of course, non-classical relationships in which at least one of the entities involved is a spatial entity play a vital role in this research. For instance, the article on the University of California, Santa Barbara (UCSB) has numerous non-classical relations regarding the protests that occurred here against the Vietnam War, protests that shaped the character of the campus for decades. For instance, the link to former California Governor Ronald Reagan, *UCSB-Ronald Reagan* is best typed *imposedACurfewToReduceRiotingAt*, which is an archetypal non-classical relation. GeoSR would fail a user seeking to learn more about Ronald Reagan's influence in the South Coast area of California if it did not report this important relationship. As such, the WCG and the WT are insufficient resources for this research due to their near complete lack of or unclear understanding of non-classical relations.

4 ExploSR: Using the WAG for Semantic Relatedness

4.1 Microstructure of ExploSR

It is important to note that because of the relative unimportance of hypernymy and hyponymy in the WAG, the WAG is a novel challenge for semantic relatedness researchers. As of this writing, there are no peer-reviewed WAG-based measures available, let alone one that is optimized to allow for data exploration. As such, it was necessary to develop our own measure, ExploSR. We chose to approach the problem from the point of view of the Wikipedia editors, the people actually creating the link structure. We started by asking what it means about the relationship between a page *A* and a page *B* when a Wikipedian creates a link between the two pages. In section three, the generic semantic *type* of these links was analyzed, but to convert these into semantic relatedness values, it is necessary to assign a quantitative measure of the *strength* and *number* of these relations. Budanitsky and Hirst (2006) note that this "scaling" of a knowledge repository network used in a SR method is "a widely acknowledged problem". Indeed, this was the key challenge in designing ExploSR. Stated more simply, ExploSR must be able to assign a quantitative relatedness measure, or weight, to each edge in the WAG. To do so, it uses the following general formulas:

If $|OL_A| > C$,

$$ExploSR_A = 1 - \frac{|OL_{A \to B}|}{C + \left(1 + \log_2 |OL_A - C|\right)} \quad \textbf{(1a)}$$

Else,

$$ExploSR_A = 1 - \frac{|OL_{A \to B}|}{|OL_A|} \quad \textbf{(1b)}$$

And if $|OL_B| > C$,

$$ExploSR_B = 1 - \frac{|OL_{B \to A}|}{C + \left(1 + \log_2 |OL_B - C|\right)} \quad \textbf{(2a)}$$

Else,

$$ExploSR_B = 1 - \frac{|OL_{B \to A}|}{|OL_B|} \quad \textbf{(2b)}$$

with the final ExploSR value being,

$$ExploSR_{A \leftrightarrow B} = \frac{ExploSR_A + ExploSR_B}{2} \quad (3)$$

In these formulas, $|OL_A|$ and $|OL_B|$ represent the total number of outlinks (the *outdegree*, in graph theory terminology) of articles *A* and *B*. $|OL_{A \rightarrow B}|$ and $|OL_{B \rightarrow A}|$ signify the size of the set of outlinks from article *A* to article *B* and vice versa. *C* is a constant that is predefined and explained below. In all cases, if either the $ExploSR_A$ or $ExploSR_B$ value is less than zero, it is set to zero[3].

The motivation behind this approach to edge weighting is straightforward. Given the nature of Wikipedia, the percentage of outlinks directed from any article *A* to any article *B* and vice versa is a good measure of the importance of the relationship(s) between A and B. However, since longer articles generally have more *relationship content*, encoded as a larger number of outlinks, some additional scaling must be done. The reasoning for the logarithm-based schema is that it was determined through extensive experience with Wikipedia that, in general, long articles are split up into sections, in each of which a cluster of references to the same articles is likely to occur. In the case of an article *B* that is extremely closely related to a long article *A*, a significant sprinkling of references to *B* is expected outside of that cluster as well. For example, in the *United States* article, links to the *Democracy* article are going to be clustered in the section on politics. However, since Democracy is so vital to the United States, it is likely to be mentioned occasionally elsewhere as well. The value *C* is the expected size of a cluster of links (C = 5 in our current implementation) and the logarithmic part of the normalization methodology approximates the number of links external to the cluster ("the sprinkling"). If equations 1a and 2a were omitted in favor of 1b and 2b for all outlink values, long articles would almost always appear to contain only weak relationships.

It is important to note that ExploSR is technically a measure of semantic distance, or the lack of semantic relatedness. We have chosen to encode it in this manner for the purposes of easily incorporating it into a Dijkstra's shortest path (Dijkstra, 1959) algorithm implementation, which is described in section 4.3.

While the formula above provides our general approach, there are a few minor data set-specific modifications. For instance, links that appear in the first paragraph – almost always a *gloss*, or summary of the article content – are treated as codifications of especially strong relationships. Similarly, we take measures to handle the unique relationships present in links between articles such as *Austria* and *Geography of Austria*.

[3] This would occur if, for example, an article *B* has 500 outlinks and the number of links from article *B* to article *A* was greater than the denominator value. In other words, in equation 2a, if C = 5, $OL_{B->A}$ = 16 and OL_B = 500, then equation 2a evaluates to approximately 1 - 16/(5 + 1 + 8.951), which is less than one. The value is then set to 0.

4.2 The Missing Link Problem

While the Internet as a whole suffers greatly from link spam, the larger problem in Wikipedia is missing links (Adafre and Rijke, 2005). This, of course, has a detrimental effect on ExploSR as a missing link essentially represents a missing relation. In the context of ExploSR, there are two types of missing links, type one and type two, both of which are important issues. In the case of type one missing links, the target of the missing link is an article that is not linked elsewhere in the page. This affects whether or not a relationship between the pages in question is identified at all. Type two missing links occur when the target of the missing link is the target of another link elsewhere in the article. In other words type one missing links affect the recognition of relationships between entities, while type two missing links affect the ability of ExploSR to identify the relative importance of existing relationships. Of course, there are some type one "missing links" that represent relationships so unimportant or weak that we would prefer that these links not be "found". "Finding" these links would be essentially introducing link spam to the data set.

In an effort to avoid the link spam problem, we currently only target type two missing links with our missing link reduction approach, which has been implemented but not applied to the whole of a WAG due to computational complexity issues. That said, our missing link processor represents a rudimentary but sufficient algorithm for this proof-of-concept stage. Future work may improve this area quite a bit, possibly enhancing the system of Adafre and Rijke (2005), which presented qualitatively promising results. Simply stated, we do a text search for all forms of links that already appear on a page and code matching non-linked forms as links. A link's "forms" include the title of the target of the link, the set of "anchor texts" (Adafre and Rijke, 2005) that are used to describe that link (i.e. the link appears as "GIS" to Wikipedia readers, but the target of the link is "Geographic Information Science"), as well as the set of redirects to the link target defined globally in the Wikipedia data set.

4.3 Macrostructure of ExploSR

So far, we have described how ExploSR scales the relationship between any two linked articles *A* and *B*. But how does ExploSR work across the entire WAG? How does this apply to the spatial context of this research? These are the topics of this subsection.

At the core of ExploSR's macrostructure is an implementation of Dijkstra's shortest path algorithm (Dijkstra, 1959). The input to this algorithm (by a user or a system; see section five for more details) is a spatial or non-spatial article *A*. The algorithm then evaluates the relations between both the articles to which *A* links, as well as the articles that link to *A*, using the ExploSR measure. It continues according to Dijkstra, summing the ExploSR values along each path, either until the entire WAG has been explored or a certain stop condition has been met. While doing this, it is recording the snippets containing each of the links it encounters. In this fashion, every relationship has a *snippet path* of sorts, even for paths that are several edges long. These snippet paths are essential to data exploration because they almost always fully explain the relationship found by the algorithm, as is noted in section three.

We made several modifications to the standard Dijkstra algorithm in order to account for the Wikipedia data set and our spatially-focused application. First, a condition has been placed in the algorithm to stop processing paths when it encounters the pure temporal articles discussed in section two. This effectively prevents the recognition of all relationships through these articles. We have done this because pure temporal articles almost always have extraordinarily weak relationships encoded in both their inlinks and outlinks (Hecht et al. 2007a). Hecht et al. (2007a) describe the example that the pure temporal article *1979* is "essentially a list of events that occurred in 1979, a list that is so disparate that it includes the acquisition of home rule for Greenland and the premiere of 'Morning Edition' on the United States' National Public Radio." (p. 4) We have found that it is better to simply ignore the relatedness of Greenland and "Morning Edition" rather than use ExploSR to estimate its microscopic general value.

Second, a similar *optional* stop condition is made available for spatial articles, albeit for an entirely different reason. When the Dijkstra algorithm encounters a spatial article, the articles that link to this article and that are linked in this article will have a large degree of spatial autocorrelation. If the user wishes to mute this effect, she can enable this stop condition. Obviously, if the user inputs a spatial article to the algorithm, this condition is not applied on the input article.

While we have now answered the question regarding the application of ExploSR to the entire WAG, we have not explained how all these values are applied to a geographic reference system. The answer to this question lies in the output of the modified Dijkstra algorithm, which is the set of spatial articles encountered by the algorithm, along with the ExploSR values of these articles and their snippet paths. This can either be a size-limited set representing the top *n*-most related articles to the input article, a value-limited set containing all spatial articles with an ExploSR value of no more than *v* from the input article, or, if computational complexity is no object, the entire set of spatial articles. For instance, a user who inputs the article *Fidel Castro* to GeoSR and sets *n* to 100 will receive the 100 spatial articles with the lowest ExploSR scores from *Fidel Castro* (figure 1), along with the attribute data described above.

5 Applications

As noted in section three, spatial articles, or articles with a geographic reference system location, can act as "sample points" for the ExploSR semantic relatedness values in the real world. It is upon this ability that we envision a myriad of applications for GeoSR.

5.1 Simple Data Exploration

The most immediately obvious application of GeoSR is to use it for point-based data exploration of the knowledge contained in Wikipedia. This application has been implemented and can be seen in figures 1 and 3. Users input an entity (which must

have a corresponding Wikipedia article) into the system, and a map indicating the *n*-most related spatial articles is presented, with the articles represented as points at their geotagged locations. Users can then click on the points to view the snippet paths for the clicked spatial article (figure 2).

We have implemented this system by exporting the output of GeoSR into a shapefile and loading this data into ArcGIS[4]. The shapefile contains three columns in its attribute table: name of the spatial entity, its ExploSR value, and its snippet path (snippet path functionality not yet fully implemented). The shapefile is visualized in ArcGIS using a reverse graduated symbol schema such that lower ExploSR values result in bigger symbols. As such, the visualization represents semantic relatedness and not semantic distance. Users can engage in data exploration by using the "Identify" tool in ArcGIS to view the snippet paths (figures 1 and 2).

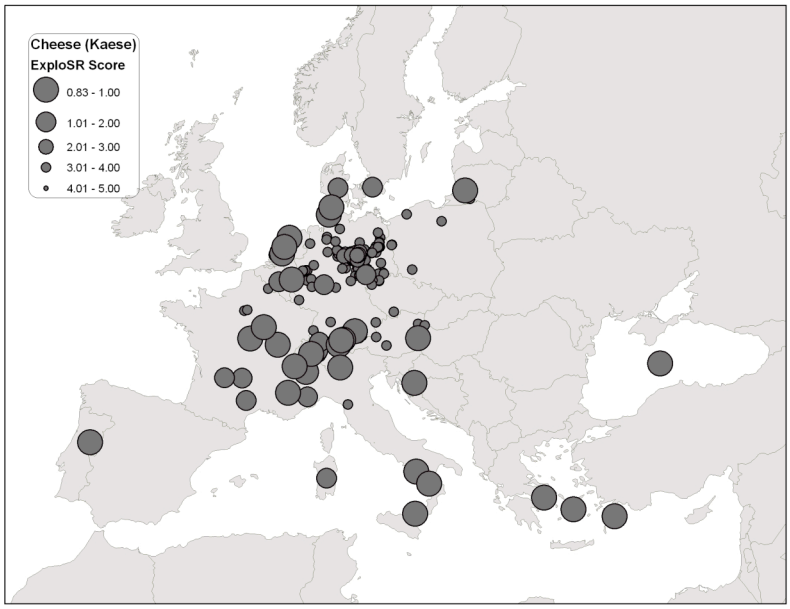

Figure 3: A visualization of the output resulting from inputting the article *Kaese* (German for *Cheese*) into the GeoSR system operating on the German Wikipedia. Spatial stemming was turned on, and missing links were not included. The top 200 locations were output, but not all are located in the region depicted above.

5.2 Area-based Query

If all spatial articles have been evaluated against all non-spatial articles (or a subset of non-spatial articles that are of interest), a user can query any extent and receive the

4 http://www.esri.com

most related non-spatial articles to that extent. This can be easily calculated using summary statistics of the SR values generated from the spatial articles located inside the chosen extent. It would also be a simple matter to explain the relatedness of these non-spatial articles to the extent using snippet paths.

5.3 Analyzing the First Law of Geography

Simply stated, the "First Law of Geography", first recognized by Tobler (1970), declares that everything is related, but nearer entities are more related than distant entities. While the nature of this "law" as actually being more of a "guideline" has been widely recognized for many years now, researchers could, by entering spatial articles as the input, have another means of exploring the degree to which this guideline holds true.

5.4 Regionalization

Many regionalization schemas and algorithms could be applied using the output of GeoSR as input. For instance, McKnight (2000) uses "basic features of homogeneity" as a means for regionalizing North America. Such uniform regionalizations could be completed by analyzing the variation in the most related non-spatial articles across space. Similarly, nodal regions could be made by evaluating the output of GeoSR when a spatial article is input.

5.5 Subsets and Algebra

While all the aforementioned applications have been explained using a single input value, there is no reason the outputs of multiple inputs cannot be combined to give new meaning to the above applications. For instance, the system described in section 5.1 could be used to examine the spatial footprint of the union of *Cheese* and *Fondue* by simply adding together the output from two iterations of GeoSR. Similarly, the applications could be used on subsets of spatial or non-spatial articles. For instance, application 5.2 could be used on the subset of non-spatial articles that are about architecture or even country music, as defined by the architecture and country music categories in the WCG.

6 Conclusions and Future Work

In this paper, we have presented two inter-linked innovations. First, we have demonstrated the benefits of visualizing semantic relatedness measures from the perspective of a geographic reference system. Second, we have created a semantic relatedness measure that is optimized for data exploration purposes. Integrating these innovations resulted in a novel data exploration environment that can form the basis for many useful applications. However, there is much work yet to be done.

First and foremost, there is no reason that GeoSR needs to be restricted to geographic reference systems. In theory, our reference system + data exploration methodologies could be applied to any *semantic* reference system (Kuhn, 2003; Kuhn and Raubal, 2003). For instance, temporal reference systems would be an easy extension as all of the above applications have simple corollaries in the temporal domain. Extending our research to semantic reference systems is the most immediate direction of future research.

Secondly, some sort of a formal evaluation is in order (we have evaluated thus far using our area knowledge of test input entities). This is a particularly difficult problem. Semantic relatedness researchers have had some difficulty evaluating their measures within their own domain, and inside the spatial domain we have the additional dilemma of the spatial dependence of opinions about relatedness between many entity pairs. Nowhere is this more evident than in the varied results of GeoSR depending on the language of the WAG. For instance, when GeoSR operates on the German WAG, no matter what its input, entities within the German-speaking world of Germany, Austria, and Switzerland always rank high, even when the input article is *Surfing* (*Wellenreiten*).

While ExploSR is currently the only WAG-based semantic relatedness measure, Zesch et al. (2007b) have expressed interest in experimenting with the WAG and surely other SR researchers will join in as well. Depending on their methodologies, it may be possible to replace ExploSR with another SR measure if that measure is proven to be higher quality and capable of producing snippet paths for data exploration. This would be another interesting area of further research.

Finally, it is our intention to analyze the extent to which relations to and from spatial entities differ from those between non-spatial entities. For instance, we would like to better investigate from a theoretical and experimental perspective why non-classical relations are so important to spatial entity relationships.

References

Adafre, S. F. and de Rijke, M. (2005). Discovering Missing Links in Wikipedia. LinkKDD (in conjunction with SIGKDD), Chicago, IL.

Albaredes, G. (1992). A New Approach: User Oriented GIS. EGIS '92.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics, *32*(1), 13-47.

Buriol, L. S., Castillo, C., Donato, D., Leonardi, S., and Millozii, S. (2006). Temporal Analysis of the Wikigraph. Proceedings of Web Intelligence, Hong Kong.

Denning, P., Horning, J., Parnas, D., and Weinstein, L. (2005). Inside Risks: Wikipedia Risks. Communications of the ACM, 48 (12).

Dijkstra, E. W. (1959). A note on two problems in connection with graphs. *Numerische Mathematik*, *1*(1), 269-271.

Elia, A. (2006). An analysis of Wikipedia digital writing. Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy.

Gabrilovich, E., Markovitch, Shaul (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. Paper presented at the Proceedings of the Twentieth Join Conference for Artificial Intelligence, Hyberabad, India.

Giles, J. (2005). Internet encyclopaedias go head to head. Nature, http://www.nature.com/nature/journal/v438/n7070/full/438900a.html

Hecht, B. (2007). Masters Thesis. Using Wikipedia as a Spatiotemporal Knowledge Repository. University of California, Santa Barbara, California, United States.

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

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

Kuhn, W. (2003). Semantic reference systems. International Journal of Geographical Information Science, *17*(5), 405-409.

Kuhn, W. and Raubal, M. (2003). Implementing Semantic Reference Systems. AGILE 2003 - 6th AGILE Conference on Geographic Information Science, Lyon, France.

Kunze, C. Lexikalischsemantische Wortnetze. Computerlinguistik und Sprachtechnologie, *2004*, 423-431.

Lakoff, G. (1987). Women, Fire and Dangerous Things. Chicago, Illinois: University of Chicago Press.

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, *38*(11), 39-41.

McKnight, T. L. (2000). Regional Geography of the United States and Canada (3rd ed.). Prentice Hall.

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

Piff, M. (1991). Discrete Mathematics - An introduction for software engineers. Cambridge, England: Cambridge University Press.

Rainie, L. and Tancer, B. (2007). 36% of online American adults consult Wikipedia; It is particularly popular with the well-educated and college-age students. Pew Internet and American Life Project. http://www.pewinternet.org/PPF/r/212 /report_display.asp

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

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

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. Economic Geography, *46*, 234-240.

Zesch, T., Gurevych, I., and Mühlhäuser, M. (2007a). Technical Report. Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. Tuebingen, Germany.

Zesch, T., Gurevych, I., and Mühläuser, M. (2007b). Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets. NAACL-HLT.