UNIVERSITY OF CALIFORNIA

Santa Barbara

Utilizing Wikipedia as a Spatiotemporal Knowledge Repository

A thesis submitted in partial satisfaction of the
requirements for the degree Master of Arts
in Geography

by

Brent Jaron Hecht

Committee in charge:
Professor Keith Clarke, Co-Chair
Professor Martin Raubal, Co-Chair
Professor Tobias Höllerer

December 2007

Utilizing Wikipedia as a Spatiotemporal Knowledge Repository

by

Brent J. Hecht

ABSTRACT

Utilizing Wikipedia as a Spatiotemporal Knowledge Repository

by

Brent J. Hecht

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 that it is important to highlight the following characteristics: First, Wikipedia is a free encyclopedia that is produced via a collaborative effort by its contributors, a group that can be joined by any person with an Internet connection and access to the site. 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. The above facts are all relatively well-known among people who use Wikipedia, which in the United States includes 36 percent of the Internet-using population. However, what is far less understood – even within the scientific community – are the opportunities for artificial intelligence, computational linguistics, social network theory and other important areas of computer science research presented by the massive knowledge repository of ubiquitously available information that Wikipedia represents. The research presented here is the first work that explores the spatiotemporal possibilities of this knowledge repository, as well as others in the future that could that display similar characteristics. Three spatiotemporal Wikipedia research projects –

Minotour, WikEar, and GeoSR – are described in detail. In addition, a spatiotemporal framework for Wikipedia research is developed in the context of frameworks useful for the computer science fields mentioned above.

Summary: Utilizing Wikipedia as a Spatiotemporal Knowledge Repository

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 that it is important to highlight the following characteristics: First, Wikipedia is a free encyclopedia that is produced via a collaborative effort by its contributors, a group that can be joined by any person with an Internet connection and access to the site. 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 have over 100,000 articles and the largest Wikipedia, English, has over 2.05 million (Wikimedia Foundation 2007). Finally, many researchers argue that Wikipedia "has probably become the largest collection of freely available knowledge" (Zesch et al. 2007a, p.1). Nowhere else can you find extensive articles on Tim McGraw, Dijkstra's shortest path algorithm, and the burrito, all in the same data set (and all items that greatly assisted in this research).

The above facts are all relatively well-known among people who use Wikipedia, which in the United States includes 36 percent of the Internet-using population (Rainie and Tancer 2007). However, what is far less understood – even within the scientific community – are the opportunities for artificial intelligence,

computational linguistics, social network theory and other important areas of computer science research presented by the massive knowledge repository of ubiquitously available information that Wikipedia represents (Gabrilovich and Markovitch 2006). The research presented here is the first work that explores the spatiotemporal possibilities of this knowledge repository, as well as others in the future that could that display similar characteristics.

This thesis is divided into four chapters. The first chapter is designed to introduce the reader to Wikipedia, outlining the data embedded in Wikipedia that are relevant to this thesis. The chapter also provides a spatiotemporal view of *Wikipedia mining* (Zesch et al. 2007a), which Zesch et al. describe as the practice of extracting the data in Wikipedia that is not in immediately machine readable form. Additionally, the chapter describes the methodology that was developed to work with the the massive Wikipedia data set in an organized and structured fashion. Chapter two presents the *Minotour* project, which generates spatial narrative-based tours from Wikipedia. In chapter three, the reader will find a description of *WikEye*, which allows users to browse the implicit structures of Wikipedia using advanced cell phone-based paper map navigation. The fourth chapter covers the use of Wikipedia to create the first-ever geographic semantic relatedness system, *GeoSR*. Finally, the thesis concludes with a summary of work completed and the work that lies ahead.

Chapter 1 – The Dataset: Wikipedia

1.1 What is Wikipedia?

It is an assumption of several algorithms described in this thesis and of Zesch et al. (2007) that the first paragraph in nearly all Wikipedia articles is a *gloss*, or a succinct definition of the article. As such, the question "What is Wikipedia?" is probably best summed up by the gloss of the English Wikipedia article on Wikipedia itself:

> "Wikipedia is a multilingual, web-based, free content encyclopedia project, operated by the Wikimedia Foundation, a non-profit organization." (Wikipedia 2007 "Wikipedia")

The article continues:

> "[Wikipedia] has been written collaboratively by volunteers around the world and the vast majority of its articles can be edited by anyone with access to the Internet." (Wikipedia 2007 "Wikipedia")

Immediately after describing the basics of Wikipedia, most authors in the Wikipedia research community then provide a brief summary of the scale and scope of the encyclopedia, probably because the scale and scope are so shocking after finding out that Wikipedia is entirely user-contributed, multi-lingual, and free. This author shall follow suit, with all statistics coming from the Wikimedia Foundation (2007). As of October 17, 2007, the last date for which full multilingual statistics were available, Wikipedia contained a whopping total of more than 8.78 million articles in over 250 languages. By far the largest Wikipedia is the English Wikipedia.

The German (0.59 million), French (0.50 million), Polish (0.38 million), and Japanese (0.37 million) round out the top five, but there are 14 languages total with over 100,000 articles. Interestingly, some of these languages are spoken by relatively few people, such as Norwegian (0.11 million articles) and Finnish (0.12 million articles). On the other hand, the Chinese and Arabic Wikipedias, for instance, have relatively small article sets, measuring in at 0.13 million articles and just 31,000+ articles, respectively, at the time of the sample. Why certain languages have extremely high article/number of speaker ratios while others have extremely low such ratios is an important open geographic question. However, the simple fact that there are a large number of languages with significant encyclopedic knowledge in their Wikipedias adds a unique dimension to using Wikipedia as a knowledge repository, a dimension that will be exploited heavily in this thesis.

1.1.1 Wikipedians: The Constructors and Maintainers of the Knowledge Repository

Who are these people who have made Wikipedia the grandiose knowledge repository that it is today? How many are there? How do they go about editing the encyclopedia? What are their motivations? Answering these and other similar questions completely is a critical component of this thesis – even though it is far from the thesis' focus – as it is important to understand who is behind Wikipedia before using it as a primary data set.

Probably the most important macro-statistic about contributions to Wikipedia is that the distribution of number of edits per Wikipedia user approximates a power-

law distribution (Voss 2005). In other words, a relatively small group of users is responsible for a "disproportionate" amount of content. It is this group that in large part defines all the Wikipedia structures defined in section 1.2, structures that are essential to all projects in this thesis. If we follow Zachte (2007) and define "Very Active" Wikipedians as those that contribute more than 100 edits per month, there were only 4,330 such Wikipedians as of October 2006.

So what motivates the small group of active Wikipedians? Several attempts to answer this question have been made in both academia and the popular press. Qualitatively, Forte and Bruckman (2005) found that the incentives that motivate this Wikipedian group somewhat resembles those that motivate the world of academia. Riehle (2006) investigates "why Wikipedia works" by interviewing three prominent administrators of the international Wikipedia community and discovers that idealism and support for the principles of Wikipedia are a large motivator. Mehegan (2006) agreed that "doing good for society" is the primary motivator, at least in the context of the English Wikipedia.

1.1.2 Accuracy of Wikipedia and Possible Use of Other Data Sets

Even more important as context to this thesis than the contributors to Wikipedia are their contributions themselves. Significant concerns have been raised both in academia, summarized well by Denning et al. (2005), and in the popular media, comically shown by Colbert (2006), about the risks of using Wikipedia data for research purposes. Critically, this research significantly reduces these risks by not

relying heavily on the *content* of Wikipedia, but rather on the *data structures* present in the encyclopedia. While the real-world excitement and practical value of the projects in this thesis are enhanced by the assumption that Wikipedia encodes a large amount of accurate (in both recall and precision) world knowledge, the validity of this research would remain intact if Wikipedia were proven to represent a blatantly non-"neutral point of view" (NPOV). For instance, if this were the case, the GeoSR project would provide a data exploration for biased world knowledge not true world knowledge, but it would still provide just as good of a data exploration environment.

In no way is the independence of the projects in this thesis from the problems of Wikipedia more evident than in their possible application on similar data sets. Of course, in theory, each project would work with any data set with the properties required by the project. In practice, two data sets are on the verge of becoming simple alternatives to Wikipedia for *all* the projects in this thesis thanks to their use of the MediaWiki[1] software, for which the software written for this research is customized. When these two data sets, Citizendium[2] and Conservapedia[3], become large enough and begin providing database dumps (see the next section), it would be an extremely simple matter to input their dumps into all the projects described here. In the case of Citizendium, which is an attempt to provide some expert oversight to the Wikipedia contribution model, doing so may provide a *more* accurate

---

[1] http://www.mediawiki.org

2 http://www.citizendium.org

3 http://www.conservapedia.com/

representation of world knowledge. In the case of Conservapedia, doing so would provide access to a knowledge repository that encodes the American conservative's world view. It would be very interesting to compare the results of all three projects amongst all three data sets.

Independence of this thesis from the Wikipedia data set aside, because of the aforementioned practical value of Wikipedia as a decently accurate representation of world knowledge, a brief overview of the Wikipedia accuracy debate is in order. The most important work in the area of Wikipedia data quality is the *Nature* investigation conducted by Jim Giles (Giles 2005). In the study, which compared the accuracy of Wikipedia with its most significant traditional rival, Encyclopedia Britannica, Giles found that both encyclopedias had the exact same number of serious errors. However, Wikipedia had a slightly larger count of minor problems, with the number of "factual errors, omissions, or misleading statements" in surveyed articles numbering 162 compared to Britannica's 123. In a separate analysis, Denning et al. (2005) writes that, "regardless of which side you're on, relying on Wikipedia presents numerous risks." They list these risks as concerns over accuracy, motives, uncertain expertise, volatility, coverage, and sources, concluding that Wikipedia "cannot attain the status of a true encyclopedia without more formal content-inclusions and expert review procedures".

1.1.3. Data dumps

Because Wikipedia is so large, it is impractical to use crawling techniques to gather data. The practice is also banned by the Wikimedia Foundation for traffic reasons. Fortunately, the foundation provides a regularly updated set of "database backup dumps" to the public. These "dumps" come in the form of XML files. Several versions of the dumps are available. In this thesis, the dump for each Wikipedia with only the most current version of each article is used. A dump of the entire history of every article is also available. Even without the full history, however, the dumps can get quite enormous. The English Wikipedia XML dump file from October 23, 2007, which is used as the data source for the English Wikipedia for much of the work in this thesis, is a gigantic 12.7GB. The October 10, 2007 German Wikipedia XML dump file (current versions only), which is also heavily used in the work presented here, is 3.63GB. Each XML dump file contains some rudimentary information about every page in the Wikipedia (title, id, etc.) along with a copy of the Wiki markup for that page.

1.2 Non-spatiotemporal Structures in Wikipedia Useful to Knowledge Repository Work

Zesch et al. (2007a) lays out the first detailed description of the sources of lexical information within Wikipedia. However, there are a few holes in their work, and, of course, they did not include anything about spatiotemporal data. As such, table 1.2a provides a major adaptation and expansion of the table presented in Zesch et al. (2007a) that highlights the structures used by some or all of the projects in this

thesis as well as those structures that are more peripherally important. The rest of this section will discuss the significant non-spatiotemporal structures in greater detail. The spatial-temporal structures are analyzed later in this chapter.

Data Source	Description of Lexical and/or Spatial Usefulness
Article page text	also referred to as Wikipedia Text (WT)
Title hierarchy	short definition of context, identification of pure temporal articles
Snippets	paragraphs within the full-text of the article
First snippet	gloss, or short definition, of subject of Wikipedia article
Temporal references	pointers to the temporal reference system
Redirects	information on synonyms, spelling variations, common misspellings, common case variations
Wikipedia Article Graph (WAG)	
Links	backbone of graph between articles
Link label	information on synonyms, spelling variations, related terms
Wikipedia Category Graph (WCG)	
Category	contains category links, sometimes short descriptions of subject of category
Category links	backbone of "folksonomy" between categories, contains almost entirely "isA" and "hasA" relations
Category memberships	locates articles within a category "folksonomy" (VanderWal 2004)
Disambiguation pages	
Disambiguation links	sense inventory
Templates	
Spatial references	point references to the WGS 1984 geographic coordinate system
Other templates	context-appropriate structured information

Table 1.2a: An expanded and altered version of a similar table in Zesch et. al (2007) that describes structures in Wikipedia useful for using the encyclopedia as a knowledge repository.

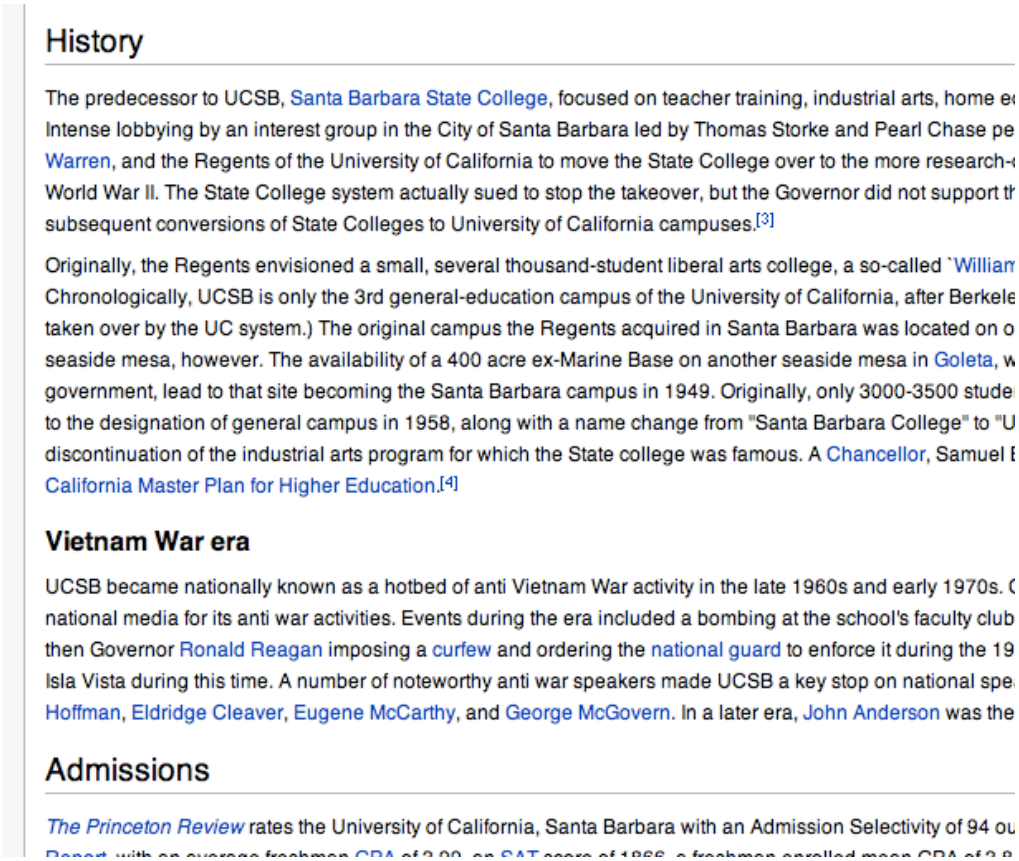
1.2.1 Wikipedia Text (WT)

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 natural text restriction prevents Wiki markup syntax - the markup language used on Wikipedia - from being

included as well. In other words, the WT is all the "normal" raw text that appears on a served Wikipedia HTML page.

The WT contains many valuable substructures, as shown in table 1.2a, of which the title hierarchy (figure 1.2b) is probably the most initially prominent. Each article has a title hierarchy, which is a structure that has not yet been explicitly identified or utilized in the Wikipedia research community, although Liu and Birnbaum (2006) made extensive use of a similar structure in the Open Directory Project[4]. The hierarchy has as its root the title of the article in question. The children of the root and their children are defined by the nested hierarchy markup contained in the Wiki markup language. Currently, the maximum depth of the tree is six (title + 5 nested subheadings). Each node of the hierarchy (including the root) can have none, one, or many snippets (see below) as leaves. The explicit storing of the location of snippets in the title hierarchy has proven quite useful for certain tasks, as is discussed below.

4 http://www.dmoz.org/



History

The predecessor to UCSB, Santa Barbara State College, focused on teacher training, industrial arts, home e
Intense lobbying by an interest group in the City of Santa Barbara led by Thomas Storke and Pearl Chase pe
Warren, and the Regents of the University of California to move the State College over to the more research-
World War II. The State College system actually sued to stop the takeover, but the Governor did not support th
subsequent conversions of State Colleges to University of California campuses.[3]

Originally, the Regents envisioned a small, several thousand-student liberal arts college, a so-called `Willian
Chronologically, UCSB is only the 3rd general-education campus of the University of California, after Berkele
taken over by the UC system.) The original campus the Regents acquired in Santa Barbara was located on o
seaside mesa, however. The availability of a 400 acre ex-Marine Base on another seaside mesa in Goleta, w
government, lead to that site becoming the Santa Barbara campus in 1949. Originally, only 3000-3500 stude
to the designation of general campus in 1958, along with a name change from "Santa Barbara College" to "U
discontinuation of the industrial arts program for which the State college was famous. A Chancellor, Samuel E
California Master Plan for Higher Education.[4]

Vietnam War era

UCSB became nationally known as a hotbed of anti Vietnam War activity in the late 1960s and early 1970s. C
national media for its anti war activities. Events during the era included a bombing at the school's faculty club
then Governor Ronald Reagan imposing a curfew and ordering the national guard to enforce it during the 19
Isla Vista during this time. A number of noteworthy anti war speakers made UCSB a key stop on national spe
Hoffman, Eldridge Cleaver, Eugene McCarthy, and George McGovern. In a later era, John Anderson was the

Admissions

The Princeton Review rates the University of California, Santa Barbara with an Admission Selectivity of 94 ou

Figure 1.2b: Two tiers of the title hierarchy are seen here in a screenshot from the UCSB page taken October 29, 2007. The first snippet belongs to the "UCSB -> History" node. The second snippet is under the "UCSB ->History -> Vietnam War era" node. Finally, the third paragraph snippet to the "UCSB ->Admissions" node.

While the title hierarchy may be more prominent, the snippet structure (figure 1.2b) is probably the most important substructure of the WT, at least in the context of this research. First identified in (Hecht et. al 2007a), 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 the title hierarchy 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 research project included in this thesis. We have found that the only context necessary for fully understanding nearly all snippets is the title of the Wikipedia article in which they appear. Most of the remaining snippets can be fully framed by providing the full location in the title hierarchy at which the snippet occurs.

Thus far, two 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 due to the encyclopedic nature of the writing style in Wikipedia, which is the second identified cause of the independence of snippets. Elia (2006) defines this style as "WikiLanguage", or "the formal, neutral and impersonal language used in the official encyclopedia articles". In other words, Wikipedians do not seek to create prose that flows from paragraph to paragraph; they seek to inform about facts in an organized, utilitarian fashion.

The first snippet in nearly all articles plays a unique role. It is defined in the style guide of Wikipedia to contain a gloss, or short description or summary, of the content of the article. While all rules in the guide are effectively no more than guidelines due to the democratic nature of Wikipedia content, the Wikipedia community has enforced this first-paragraph-as-gloss convention relatively strictly. An easily obtainable gloss about an article's subject allows for the adoption of many computational linguistics techniques from WordNet (Miller 1995), a very prominent knowledge repository in that discipline. These glosses are also a critical part of the Minotour project, as is described in chapter three.

While the research here 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. Research in this area with be discussed in chapter four.

1.2.2 Wikipedia Article Graph (WAG)

Aside from the snippet construct, the Wikipedia Article Graph (WAG) is the single most important structure within Wikipedia in the context of this research. 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 $G = (V, E, \varphi)$. V is the set of vertices in the graph, E is the set of edges, and φ 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 φ 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 the research projects in this thesis, which were generated in October 2007, the English Wikipedia had $|A| \sim= 2.05$ million and $|L| \sim=$ 45 million and the German Wikipedia had had $|A| \sim= 0.69$ million and $|L| \sim= 15.0$ million. The size of the graph creates certain challenges and forces long processing times, issues that will be discussed later in this section. 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 will be discussed in full detail in chapter four.

Wikipedians can create a link between two articles in the WAG by simply enclosing text in two [[brackets]]. The word "brackets" in this case would appear as a link to article on "brackets". If the word "brackets" is contained in the redirect table (perhaps to the word "bracket"), then the user will be immediately forwarded to the "bracket" article. Providing different labels for links is also possible. The expression [[big fat paper|thesis]] would appear as a "big fat paper"-labeled link to the user, but would point to the "thesis" article.

Buriol et al. (2006) provide an overview of the (then-) current state of the English WAG as well as its temporal properties. They found that the WAG is a scale-free graph with an indegree (number of links) distribution that follows a power-law. In other words, the probability that an article has *x* number of inlinks is proportional to 1/(γ^*x*). They add that γ is approximately the same as that of the Web, and that it has remained very constant despite the meteoric growth of Wikipedia during their study period (1/2002 - 4/2006). Finally, they also point out that the WAG has been becoming denser and denser, with the average number of inlinks increasing from seven to 16. In the October 2007 English Wikipedia we found this to have increased significantly to 22.6. The October 2007 German WAG was surprisingly similar in density, with the average number of inlinks equal to 22.7.

1.2.3 Wikipedia Category Graph (WCG)

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 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.

Importantly, Strube and Ponzetto (2006) note that folksonomies, unlike carefully defined ontologies, strive for excellence via collective approximation rather than intensive scientific rigor. As a result, unlike true taxonomies, the WCG is a graph that describes more than just hypernymy and hyponymy, or *is-a* relationships. At least meronymy, or the *has-a* relationship, is also represented (Strube and Ponzetto 2006) in small part. This heterogeneity of relationship types is what has led some SR researchers to identify the WCGs as good candidates for evaluating semantic relatedness measures. However, compared to the WAGs, the WCGs display an incredibly small amount of relationship type diversity, a fact which is discussed in full detail in chapter four.

Zesch and Gurevych (2007b) extend the work of Zlatic et al. (2006) to the German WCG by confirming that the German WCG is a scale-free, small world graph and thus similar to other lexical semantic networks such as the WAG, WordNet and Roget's Thesaurus.

1.3 Spatiotemporal Structures in Wikipedia Useful to Knowledge Repository Work

As of the publication of this thesis, no geographer has examined Wikipedia, let alone investigated its use as a knowledge repository. As such, it is important here to establish an introductory framework that can be used to view Wikipedia from the perspective of the Geographer. The framework is summed up as follows: examined from a spatiotemporal viewpoint, Wikipedia articles can be described as having or not having any combination of these two key properties: spatial reference(s), and/or temporal reference(s). This chapter will describe this framework in greater detail, as well as expand upon its benefits and limitations. Attention is also paid to the methods used to make explicit these sometimes-implicit characteristics of Wikipedia. This section is meant as a spatiotemporal extension to the outline of Wikipedia defined in the preceding section.

1.3.1 Spatially Referenced Articles

Wikipedia articles can have no spatial references, one spatial reference, or two or more spatial references. Articles without a spatial reference are termed *non-spatial articles*, whereas articles with one spatial reference of a specific type to be described

in this section are termed *spatial articles*. Currently, all projects described in this thesis do not support multiple spatial references.

1.3.1.1 The Spatial Referencing Process

Wikipedia articles are spatially referenced by Wikipedia users through a geotagging process. In the Web 2.0 world, geotagging usually involves the (too) simple task of "tagging", or attaching, a latitude and longitude coordinate pair to a web document. 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}}") and essentially describe a function name and its parameters. The function is executed upon the loading of the Wikipedia article by the browser of a visitor to Wikipedia. For instance, a commonly used template in the English Wikipedia is the *main* template, whose syntax and output can be seen in figure 1.3a.

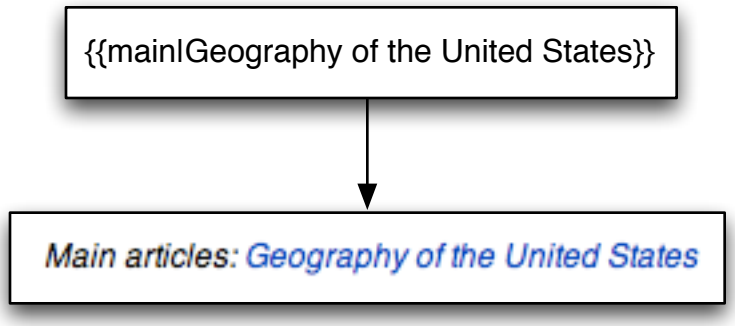


Figure 1.3a: An example of the *main* template Wikiscript, as well as the output of the *main* template seen by Wikipedia readers.

Before Google began including a parsed layer of spatial Wikipedia articles in its Google Earth software, there were dozens of templates used by Wikipedians to attach spatial references to Wikipedia articles. In November 2006, a total of 127 were counted. Each of these templates has slightly different features, most of which involve linking to scale- and location-specified version of Google Maps, Yahoo! Maps, Terraserver images, and other third party mapping sites. The massive syntactical variety of these templates despite their high semantic similarity highlights an important challenge of using Wikipedia as a geospatial knowledge repository: in addition to Wikipedia's democratic *content*, it also has democratically-defined *structure*. As a result, it can be quite difficult to cast many aspects of Wikipedia's structure into a more computer-usable structure. This is a theme that runs through this entire thesis, no matter the specific project.

Since Google arrived on the scene, Wikipedians have made a concerted effort to standardize spatial references around the *coord* template, a portion of whose syntax is described in Figure 1.3b. Nevertheless, many of the deprecated templates are still in wide use.

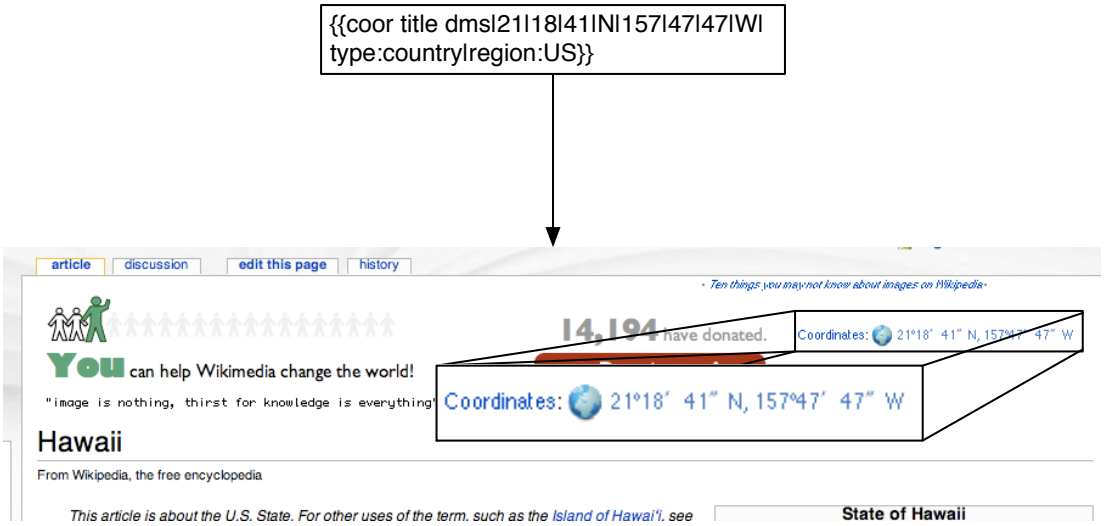


Figure 1.3b: A coord Wikiscript syntax example (above) and the output of the sample below. The portion of the top of the Hawaii page that is the output is magnified for effect. Note that the large-area region of Hawaii is geotagged to a single point. This scale issue is discussed in detail later in this chapter.

The situation in the German Wikipedia is much more standardized. As of April 2007, the German Wikipedia had nearly all of its georeferences in the format of *Koordinate Artikel* and *Koordinate Text* templates.

1.3.1.2 Single Spatial References and Multiple Spatial References

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 we term spatial articles. However, the same templates (with different parameters) or very slightly modified templates are also used

quite often within the body of a snippet 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. For instance, as of June 2007, the article on George Orwell contained a spatial reference to the town in which he wrote most of 1984. This spatial reference is hardly enough to consider the article on George Orwell to be a *spatial article*.

However, many times these inline references provide a way for users to implicitly provide multiple spatial references that apply to the entire article, as this functionality is currently not provided by any template in Wikipedia. An example of this type of spatial referencing occurs in the "Extreme points of Brazil" article, in which the northernmost, southernmost, etc. points of Brazil are described using spatial reference templates, all of which point to valid spatial references for the article. However, as a result of the inherent ambiguity of inline spatial references, multiple spatial references for articles is not supported in the systems described in this thesis.

1.3.1.3 Degree of Geotagging Coverage and Resulting Problems

The degree of geotagging coverage of English Wikipedia articles is far from extensive. In other words, there are many articles in Wikipedia that should have geotags that do not, thus preventing our systems from identifying them as the spatial articles they are. In an informal study conducted in November 2006, it was found that 20 to 30 percent of English Wikipedia articles (24.9 percent +/- 4.6% at a 95%

confidence interval) should be referenced to a single location, where this condition was determined subjectively. As of late May 2007, only around 97,500 (Kühn 2007) single-referenced geotagged articles exist, out of a total of around 1.1 million regular articles (~8.8 percent). Interestingly, the German Wikipedia fares quite a bit better, with about 25 percent of its articles geotagged (Kühn 2007)

Because geotagging, like everything else in Wikipedia, is done collaboratively by Wikipedians, coverage varies across different areas of the encyclopedia. For instance, as of late July 2007, only 9.0 percent of the 676 articles listed under the category "California High Schools" were geotagged, while 23 percent of the 221 articles listed under the category "Amusement Parks in the United States" had solitary spatial references. Note that both categories should have close to a 100 percent geotagging rate, as nearly every member of each is easily georeferenced to a single location. On the other hand, 216 of the 219 (98.6 percent) of all articles in the "Airports in California" category were tagged with coordinates, meaning only 1.4 percent of this category's articles have missing coordinates.

When mining for spatially referenced articles, recall is entirely dependent on geotagging coverage, unless external gazetteers are used. In other words, without outside help that can make explicit the implicit spatial references within a Wikipedia article title, articles cannot be identified as spatially-referenced unless they are geotagged. With at least around 50 percent of spatial articles not tagged as such in the English Wikipedia, all algorithms that depend on identifying spatial articles will be inherently inaccurate, a fact that affects all systems described in this thesis.

However, the extent of coverage of the German Wikipedia provides some hope that geotagging in the English Wikipedia will improve quickly.

1.3.1.4 Severe Scale Problems

Despite the fact that there is nothing to stop a Wikipedian from writing a template to support polyline or polygon-based geotagging, it is likely that no such template exists and certainly no such template has come into anything resembling wide use. This has left the Wikipedia community with a severe scale problem. Namely, standard cartographc practice suggests that the appropriate scale for using the set of Wikipedia spatial references is one in which the features that are referenced would appear more-or-less as points. With features as large as the state of Alaska represented as a point, the data is only technically valid for use on a scale similar to that of a picture of the Earth taken from beyond the Moon. However, these projects are all presented as proof-of-concept works, and the number of spatial references with drastic scale problems are minimal enough that they could be solved using manual correction or basic georeferencing techniques (in combination with an external dataset of polygon or line information).

1.3.2 Temporally-referenced Articles

When viewed from the temporal point of view, Wikipedia articles break down into three groups: articles without temporal references, articles with temporal references, and articles about purely temporal phenomena. This section will first

discuss this last class of articles, as they provide the anchor by which temporal references operate. Next, temporal references in articles will be explained and analyzed.

1.3.2.1 "Pure Temporal" Articles

Pure temporal articles are articles that describe a temporal entity. Examples include the articles titled "1976", "the 19th century", "1960s", "April 29". These articles are unique not only for their subject matter, however. They are also special in that they are one of the only classes of articles in which all outlinks and inlinks display a near uniformity in type; all links to and from these articles tend to be "happened on/in" links.

These articles are often a nuisance to those who wish to use Wikipedia as a spatiotemporal knowledge repository, as is explained in more detail in later chapters. However, these articles are essential to temporal referencing: many of them provide anchors on the temporal coordinate system to which all explicit temporal references point. For instance, articles that link to the article "1976" are referenced to a timeline in the same way that coordinates provide a reference to a geographic reference system. While not used explicitly in any of these projects, this temporal referencing will be important to future work. Of course, not all pure temporal articles provide good temporal references. The article on "October 29", of course, represents a highly ambiguous reference with much less utility.

1.3.2.2 Temporal references

As was the case with spatial referencing, Wikipedians have created an elaborate supply of *explicit* and *implicit* temporal references. However, unlike with spatial references, our systems are able to take advantage of a significant portion of the implicit references.

Explicit temporal references in Wikipedia, as is noted above, take the form of links to pure temporal articles. Using the standard Wikipedia link syntax, a Wikipedian can provide temporal references to sentences relatively easily. Explicit temporal references are a unique benefit to Wikipedia, and solve for a significant percentage of Wikipedia articles the fixed-format (i.e "<month> dd, dddd" or <dd>th Century", where d are digits with certain properties) temporal expression disambiguation problems discussed in (McKay and Cunningham 2000). However, not all fixed format date expressions are linked to pure temporal articles, and McKay and Cunningham's work can be applied to parse out with high overall accuracy the remainder of fixed-format temporal references (the implicit references).

Although not examined in detail in this thesis, it is possible to apply more advanced temporal expression analysis to recognize and resolve the non-fixed format temporal expressions in Wikipedia such as "the next year", "the next week", as is discussed in (Mani and Wilson 2000) and (Mani 2004). However, due to Wikipedia's encyclopedic tone and the fact that, on average, at least seven authors have contributed to each Wikipedia article (as is discussed above), a larger percentage of temporal references are in a fixed-format form as compared to a narrative text, for

example. As such, given the limited accuracy of some of the more advanced methodologies, the cost/benefit ratio of using advanced temporal information extraction is likely quite high.

1.4 Dataset Preprocessing - WikAPIdia

In section one of this chapter it is noted that Wikipedia data is obtained via XML database dumps provided by the Wikimedia Foundation. Processing the explicit and implicit information in these dump files from XML into a usable set of structured database tables and writing an application programming interface (API) to access this structured data represented a significant portion of the everyday work done for this thesis. Because the use of Wikipedia as a knowledge repository is very new, no suitable freely available code was available for parsing the dumps or providing structured accessing to the dumps when work began in October 2006. As such, I designed my own Wikipedia XML dump file parser and Wikipedia API called *WikAPIdia* that meets the needs of each of the projects in this thesis. WikAPIdia is written entirely in Java and uses MySQL as its storage back-end. Extensive use is made of Java's regular expression functionality and MySQL's JDBC plug-in.

Since October 2006, two other researchers have made similar work available. *Wikipedia Preprocessor*, or *WikiPrep*, (Gabrilovich 2007) is written in PERL and produces text files of explicit and implicit information contained with the XML dumps. It is targeted toward the AI, information retrieval, and related fields. It does not, however, support multiple languages, The *Java Wikipedia Library*, or *JWPL*

(Zesch et al. 2007a), uses a similar Java-MySQL architecture as WikAPIdia and also focuses on making explicit and easily-accessible many implicit Wikipedia structures. In fact, the general system architecture of WikAPIdia and JWPL are nearly identical. JWPL supports several different Wikipedia languages.

WikAPIdia compares well against these both projects and should be a boon to the Wikipedia community once it is made available. For instance, WikAPIdia displays all the benefits (and more) of JWPL listed by Zesch et al. (2007a): "(1) decoupling the API implementation from changes in the MediaWiki software, (2) making research results reproducible, (3) explicitly storing the information, scattered in the original database structure, and (4) computational efficiency for large natural language processing tasks". WikAPIdia should be especially useful for researchers in the narrative intelligence and geography communities, as it provides special access to the explicit and implicit structures in Wikipedia that assist in its use in narrative and spatiotemporal applications. It also has extensive built-in graph mining capabilities, which should be appealing to researchers. While many of the relevant details of WikAPIdia are discussed in each of the following chapters, the rest of this section will briefly summarize the more general algorithms and data structures in WikAPIdia.

1.4.1 Multilingual (Multi-Wikipedia) support

WikAPIdia is designed from the bottom-up to support Wikipedias of any language. At the time of this writing, the English, German, and Spanish Wikipedias are supported, but it would be trivial matter for a native speaker to add support for

additional Wikipedias. In order to do so, it is only necessary to write several regular expressions and a couple of simple Java methods. Once these are specified (in the form of a Java subclass), all importation and wrapper functionality will be supported.

1.4.2 Database dump XML file importation

WikAPIdia takes the single Wikipedia database dump XML file as input and stores data into 15 different MySQL tables, some of which are used by all projects in this thesis and others of which are for a specific project. Figure 1.4a shows these some of these tables with a description and the number of rows in each for several of the major Wikipedias.

Table Name	Fields	English Rows (10/18/07)	German Rows (10/10/07)
articles	id, title, inlinks, outlinks, bytes, spatiallyReferenced, isPureTemporal, error	2,082,718	659,995
links	startID, endID, startTitle, endTitle, snippetNumber, isSubarticle, isSeeAlso, isMissingLink, headerLevel, isRedirect	47,319,388	15,006,931
snippets	id, snippet, snippetNumber, header, length	41,607,528	14,946,702
redirects	aliasID, aliasTitle, synsetID, synsetTitle, toDisambiguation	2,135,743	461,580
categoryGraph	childID, parentID, childName, parentName	538,062	82,584

Figure 1.4a: Descriptions of five tables in the WikAPIdia MySQL databases, along with row counts for each table for the English and German Wikipedias.

There are several large challenges inherent to parsing Wikipedia. First, some of the Wikipedia markup language syntax can be a challenge to process. While most of the syntax can be handled with simple regular expressions, other grammatical constructs require state to understand, particularly Wikipedia templates and Wikipedia images, both of which can be nested. As such, portions of WikAPIdia behaves like a rudimentary programming language compiler. If strict adherence to correct syntax were required on Wikipedia, WikAPIdia would operate without errors. However, the second major challenge to parsing Wikipedia is that a small percentage of Wikipedia articles and pages are saved with incorrect syntax. Such is the effect of having such a large population use a markup language without strict syntax controls. For instance, sometimes a missing closing "}}" will often occur in templates. Other times users will forget to close an image link. WikAPIdia does its best to recover from such errors, but inevitably certain articles will remain unparsed. For instance, 5,784 parsing errors were encountered out of approximately 5,365,000 pages (articles, redirects, categories, etc., of which 1.916 million were articles) in the July 17, 2007 English Wikipedia, making the error rate approximately 0.1 percent.

While the data is being imported, a significant amount of structural processing is being done using pre-loaded indices. First, redirects must be "resolved". Redirects are aliases that are used by Wikipedians to prevent multiple articles about the same topic cropping up under different names as well as for navigational purposes. For example, there exists a redirect from "UCSB" to "University of California, Santa Barbara". During this resolution process, each entry in the links

table that points to redirects is "redirected" to point to the corresponding full article. Next, inlinks must be counted and article IDs must be copied into the links table. Each article in Wikipedia has a unique title and a unique ID number. A similar process must be done on the categoryGraph table in order to enable easy access to the category ID numbers of the supercategories of each category.

The time necessary to do a full import of a dump file can be quite long. WikAPIdia's speed is also highly dependent on the amount of available memory and the data transfer rate of the hard disk(s) on which the XML file and the database tables are located. On a slower computer, the English Wikipedia can take two to three days to fully import. The German Wikipedia usually finishes in about 18 hours.

1.4.3 Wrapper Design

While the details of the class structure of the API access to the structured Wikipedia data available in the parsed MySQL tables is outside the scope of this thesis, it is important to go over a couple of the basics. Most importantly, the classes in WikAPIdia are designed to structurally and functionally mimic the explicit data structures in Wikipedia. For instance, the Article class contains functions like "getLinks" and "getCategoryMemberships". Similarly, classes with graph mining and other graph processing algorithm functionality are used to model the WAG and WCG. Second, each of the projects shown here have resulted in numerous other classes that provide important, albeit peripheral, general functionality to the API. For instance, the development of the Minotour AJAX software discussed in chapter two

has resulted in a general framework for executing all operations in the API remotely from a web browser, functionality that could prove useful for demoing the two other projects. Finally, it should be noted that, WikAPIdia is a massive software package which is surely tens of thousands of lines of code long and contains over 75 Java classes.

Chapter 2 – Minotour: Generating Educational Tourism Narratives from Wikipedia

2.0 Abstract

We present a narrative theory-based approach to data mining that generates cohesive stories from a Wikipedia corpus. This approach is based on a data mining-friendly view of narrative derived from narratology, and uses a prototype mining algorithm that implements this view. Our initial test case and focus is that of field-based educational tour narrative generation, for which we have successfully implemented a proof-of-concept system called Minotour. This system operates on a client-server model, in which the server mines a Wikipedia database dump to generate narratives between any two spatial features that have associated Wikipedia articles. The server then delivers those narratives to mobile device clients.

2.1 Introduction

Heritage/cultural tourism, ecotourism, agritourism, and other types of information-centric tourism have been of increasing importance in recent years, both to the tourist and to the tourism industry. However, these types of tourism have not yet fully taken advantage of the dramatic increase in available data that has been a hallmark of the Information Age, particularly with respect to the adoption of mobile technologies (Brown and Chalmers 2003). One aspect of tourism that is particularly in need of better mobile device applications is that of educational tourism. Most mobile technologies aimed at the tourism market are either (1) tourist tools that can

do nothing more than inform users of optimal or nearby tourist facilities or attractions and thus can do little educating (Kim et al. 2004, etc.), or (2) educational devices that are limited in scope due to the amount of required custom content development (Isbister and Doyle 2003, etc.). With Minotour, we attempt to fill this vacuum, and our central methodology is the employment of intelligent narrative technology, particularly data mining techniques informed by narrative theory.

Why use narrative? Much research has concluded that humans have an inherent predilection towards narrative approaches (Mateas and Sengers 2003). In addition, aside from the innateness of narratives to the human experience, narrative has been shown to play a particularly important role within the two constituent fields of our educational tourism test platform - education and tourism - as well as in the combined platform itself. Wells (1986), Mott et. all (1999), and many others have identified the myriad benefits of narrative to education and Gretzel and Fesenmaier (2002) have done the same for tourism. In the educational tourism context, Lanegran (2005) concluded that a high-quality educational tour is one in which a cohesive story is woven while traveling through the landscape. A similar assumption is made by Isbister and Doyle (2003).

Minotour utilizes Wikipedia as its primary data source, thus eliminating the need for any knowledge base development, which, as noted above, is a rarity in the world of automated tour guides with an educational focus. Through the use of the Wikipedia corpora, Minotour inherits all of Wikipedia's user-friendly advantages, such as democratized and free information that is global in scope. We also utilize

several other unique properties of Wikipedia, as described in section three. However, the text of Minotour's generated narratives – derived directly from Wikipedia – obviously also exhibits Wikipedia's well-publicized disadvantages, most notably with regard to concerns about accuracy, motives, uncertain expertise, volatility, coverage, and sources (Denning et al. 2005).

In its current implementation, Minotour operates on a simple premise: with a mobile device, the user selects a destination and a narrative tour from the user's current location to that destination is provided. The user's position is determined via GPS and the tour is generated on a server and delivered back to the client. These tours are designed to be experienced while the user moves to the destination (in our prototype, this is done via text display; we are also exploring other options, including audio). An example tour with Berlin's Brandenburg Gate as the start feature (feature *a*) and Unter den Linden as the end feature (feature *b*) is provided in figure 2a. In this chapter, we will walk through the theory and methods that are used to generate these tours. The second section provides a quick summary of related work in narrative intelligence. In the third section, we discuss the Wikipedia context of the project and the operation of the narrative generation algorithm. In the fourth section, we outline our understanding of narrative theory and the two key narrative cues, unity and progression, that underlie our conceptualization of the ideal narrative function, the function on which the narrative generation algorithm is based. In the fifth section, we detail the implementation of the project. An extension of the Minotour project,

WikEar, is covered in section six. Finally, we close with concluding thoughts and a description of our ideas for further future work.

Brandenburg Gate
Finally, when the Berlin Wall fell in 1989, the gate symbolized freedom and the unity of the city. It re-opened on 22 December 1989 when the West German Chancellor Helmut Kohl walked through to be greeted by the East German Prime Minister, **Hans Modrow**.
Hans Modrow
After the German reunification, he served as a member of the Bundestag and currently serves as a member of the **European Parliament**.
European Parliament
The European Parliament (formerly European Parliamentary Assembly) is the parliamentary body of the European Union (EU), directly elected by EU citizens once every five years. Together with the Council of Ministers, it composes the legislative branch of the institutions of the Union. It meets in two locations: **Strausbourg** and Brussels
Strausbourg
A revolution in 1332 resulted in a broad-based city government with participation of the guilds, and Strasbourg declared itself a free republic. The Straußburger Münster was completed in 1439, and became the World's Tallest Building, surpassing the Great Pyramid of Giza. During the 1520s the city embraced the religious teachings of **Martin Luther**, whose adherents established a University of Strasbourg in the following century.
Martin Luther
A minority viewpoint disagrees with the attempt to link Luther's work causally to the rise of Nazi anti-Semitism, arguing that it is too simplistic an analysis...Johannes Wallmann, professor of church history at the **Humboldt University of Berlin**, writes that Luther's writings against the Jews were largely ignored in the 18th century and 19th centuries.
Humboldt University of Berlin
Its main building is located in the centre of Berlin at the boulevard **Unter den Linden**. The building was erected by Prince Heinrich of Prussia. Most institutes are located in the centre, around the main building, except the nature science institutes, which are located at Adlershof in the south of Berlin. The University continues to serve the German community.
Unter den Linden
Unter den Linden...is a street in the centre of Berlin, the capital of Germany. It is named for its Tilia or lime trees (also known in North America as basswood trees) which line the grassed pedestrian mall between the two carriageways. Unter den Linden runs east-west from the Brandenburg Gate in the west to the Schlossbrücke (Castle Bridge) over the River Spree in the east. The major north-south street crossing Unter den Linden is Friedrichstrasse.

Figure 2a: A sample narrative tour generated by our prototype version of Minotour. In this tour, Brandenburg Gate is spatial feature *a*, Unter den Linden is spatial feature *b*, and s = 7 (see section 2.3 for details about variables)

2.2 Related Work in Narrative Intelligence

In the narrative intelligence framework laid out by Mateas and Sengers (1999) and Mateas and Sengers (2003), the Minotour narrative generation system draws most heavily from the then-state-of-the-art (1999) body of work designed to support human narrative intelligence. Indeed, the key presumption in Minotour is that there is inherent value - increased cohesion, improved recall and comprehension, the ability to overcome certain cognitive obstacles, increased synergy with how tourism is experienced, closer alignment with geography education, etc. - in providing information to users in a form that is easier to interpret as a narrative than in the dominant manner in which Internet information is accessed. In the context in the 2003 version of Mateas and Sengers' framework, this trait of Minotour places it in the Narrative Interfaces category, as Minotour essentially provides a narrative-based interface to Wikipedia, albeit a unique and highly spatial one.

In order to provide narrative support, however, we make a concrete assumption as to the definition of narrative, at least in the context of our system. This definition, outlined in section four, is used to generate stories, thus placing our work also well within Mateas and Sengers' Storytelling Systems category of work, and within the story-centric sub-category.

While Minotour's feature base lies firmly in narrative support and Storytelling Systems, Mateas and Sengers' Interactive Fiction and Drama category also applies to Minotour. Users interact with Minotour in both spatial and non-spatial ways: they are only allowed to obtain tours beginning in their present spatial location and end

destinations are chosen via the medium of the mobile device. In addition, because Wikipedia is editable by any user, a measure of interactivity pervades all aspects of the tours. This feature of Wikipedia, along with our system's ability to bring out emergent features of the link structure and raw text of the encyclopedia, allows any user to influence the narrative that she and all other users experience.

2.3 The Narrative Generation Algorithm

In order to understand Minotour's narrative generation algorithm - the computational methodology we use to generate tours such as that in figure 2a - it is necessary to first highlight some properties of Wikipedia.

2.3.1 The Wikipedia Context

Two spatiotemporal categories of Wikipedia articles are critically important here: articles that can be referenced to a spatial entity and articles that cannot. We refer to the former type of articles as "3D articles" because they exist in both the two-dimensional space defined by the spatial reference system used in Wikipedia (World Geodetic System 84) as well as in Wikipedia space, and the latter as "non-3D articles", because they have no spatial reference. Of course, within the context of our educational tourism test bed, the start and end articles of our tours must be 3D articles. 3D articles are identified through user-embedded latitude and longitude information and a basic georeferencing process.

Secondly, it is important to point out that a key benefit of using the Wikipedia corpus is that the online encyclopedia has a large hyperlink structure between articles (the WAG, introduced in chapter one). We take advantage of this structure in every step of our algorithm. The English Wikipedia had 32.1 million links as of October 2006 (Zachte 2007), with nearly all links representing some kind of meaningful semantic relationship.

Finally, as noted in the introductory chapter, because Wikipedia is collaboratively edited and the average Wikipedia article is contributed to by at least 7 different authors (Buriol et al. 2006), the paragraphs in Wikipedia tend to be more disjoint and contain fewer references to each other than those in other plain text corpora. Recall that Wikipedia's encyclopedic writing style also contributes to this phenomenon. As such, we are able to consider paragraphs as individual entities called "snippets" and re-arrange them to our hearts content without destroying their semantic value. Because of our current mobile device delivery platform, we have further constrained the definition of valid snippets to be Wikipedia paragraphs between m and n characters in length, with m and n currently set to 200 and 600.

2.3.2 Operation of the Algorithm

Concisely, our narrative generation algorithm operates with the following goal:

> Identify the narrative n_{best} of length s from 3D article a to 3D article b, where $Q(n_{best}) = \max(Q(n))$ for all n of length s between a and b and Q is the pre-defined

"narrative evaluation function". (n is a path through Wikipedia's link structure between a and b; s is the number of snippets in the desired tour and is derived from the distance between a and b in geographic space.)

Between nearly all pairs of 3D articles for all s greater than a small number, there are dozens to thousands of possible paths through WAG, even when the path length restriction of s is implemented. We identify these paths with the algorithm shown in figure 2b. The algorithm is given as input 3D article a, 3D article b, and s.

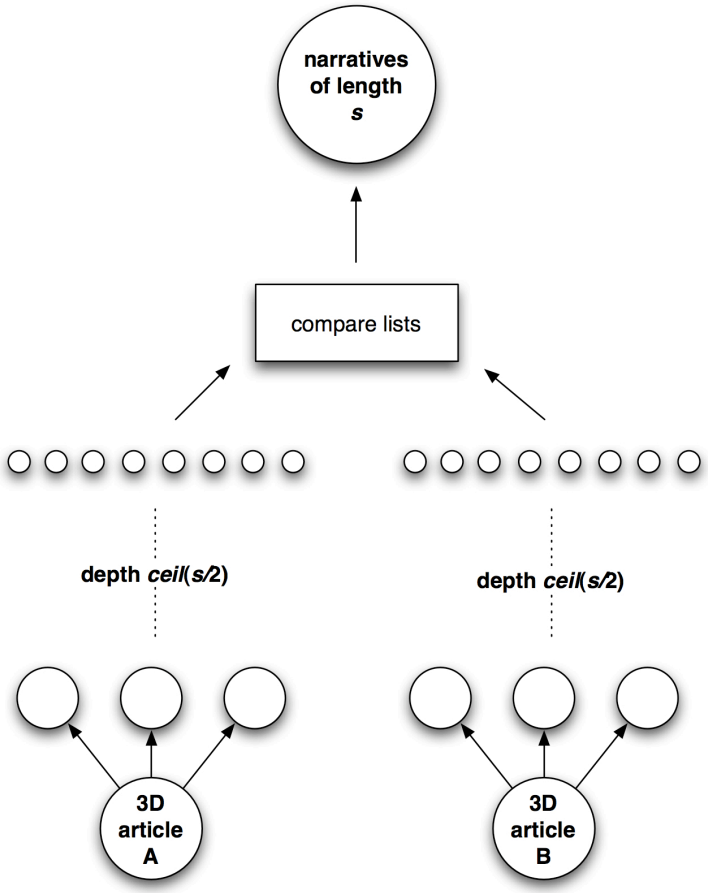


Figure 2b: A diagram depicting the operation of the path finding portion of the narrative algorithm. Once found, the narratives of length s are fed into the Q function, and the narrative with the optimal Q output is selected and delivered to the user. The above example is for the case of an odd s. The algorithm is slightly different even for s.

Conceptually, the algorithm works as follows: Starting from *a* and *b* (two separate trees), the algorithm identifies all of the out-links and in-links between *a* and *b* and other Wikipedia articles. An out-link is a link pointing from the article being considered, *x*, to a different article, *y*; an in-link is a link from *y* to *x*. We have determined that these types of links have equal semantic value for our narratives. Once the algorithm has located the destinations of the out-links and origins of the in-links, it examines the links of these destinations and origins. It continues recursively until level $ceil(s/2)$ or $ceil(s/2) + 1$ is reached, depending on whether *s* is even or odd. The lists of articles at the leaves of each tree are then compared. Any article that appears in both lists represents a connection between the trees, and thus a path between *a* and *b*. Once identified, the paths between *a* and *b* are assessed using the narrative evaluation function *Q*, and the path with the optimal output is returned. $Q(n)$ is defined as the total error (RMSE) of path *n* to a predefined ideal narrative function *Qi*, shown in figure 2c. *Qi* is two-dimensional, with snippet number on the *x*-axis and snippet host article in-links on the *y*-axis. Why have we defined the ideal narrative function in this way? The answer to this critical question lies in our novel use of narrative theory, and is the subject of section four.

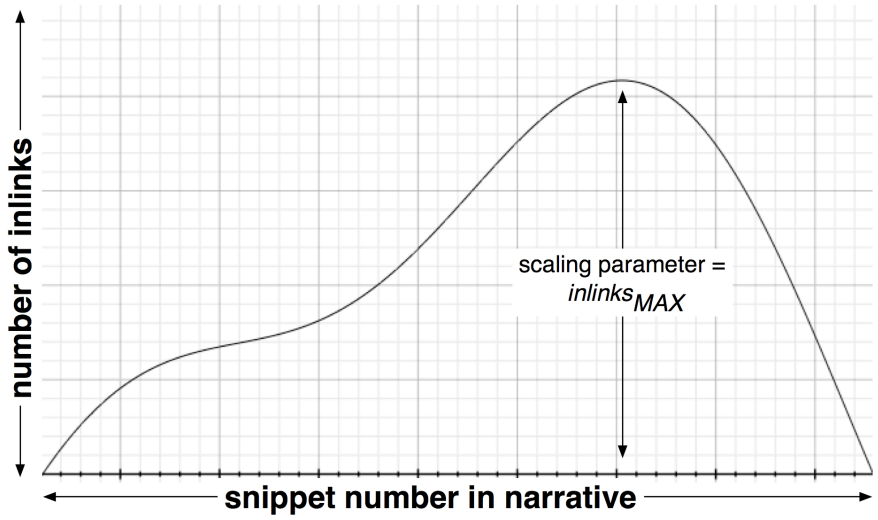


Figure 2c: The ideal narrative function, *Qi*. The narrative algorithm chooses the path *n* that most resembles this function.

While the implemented algorithm is derived from the conceptual algorithm described above, there is one key difference between the two. If the algorithm were to examine all of the possible links between each article, it would quickly become extremely slow and resource intensive. This problem is well-known in computer science and is related to each tree's "branching factor", or number of new articles to be examined per article. It is easy to see that if the first article has 900 total links (say, 800 in-links and 100 outlinks) and each of those 900 linked articles have 900 of their own links, the amount of articles to be processed becomes untenable ~(900^(*s*/2)). To solve this problem, we have turned to an effective, albeit basic and non-optimal solution that uses a simple heuristic. At each article, we only examine *BF* links, where *BF* is set to an arbitrarily low number that experimentally maximizes effectiveness while also considering the limited computational resources of our

prototype environment. We have had success with $3 \leq BF \leq 10$. Taking a hint from A* (Hart et. al 1968), we select the *BF* articles by sorting the in-links and out-links by their likelihood to produce narratives that will result in high narrative evaluation function (*Q*) values. We have found that this approach exhibits results that, while not optimal, are more than satisfactory.

While we try to keep the narrative dependent entirely on the properties of the WAG, we explicitly prevent one key group of articles from appearing in the narrative: purely temporal articles. These are disallowed from the narratives because, as is written in the introduction, purely temporal articles almost always exhibit uninteresting and non-educational semantic connections with their neighbors in Wikipedia space. For example, the 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. Other purely temporal articles include "1970s", "April 29", and "11th Millenium".

2.4 The Idealized Narrative Function

In the previous section, we show that our narrative algorithm's functionality can be simply described as trying to find the path *n* through the WAG between *a* and *b* that is most similar to an idealized narrative function, *Qi*. In this section, we discuss the two characteristics of narrative experience – unity and development - that we model in this function, as well as our reasoning for choosing these two

characteristics. We hypothesize that when these characteristics are successfully embedded in a generated Wikipedia text, the resulting textual object will be read as a narrative, and the travel from *a* to *b* will be considered a narrative experience. Unless otherwise noted, nearly all of the ideas and concepts in this section are derived from (Hecht et al. 2007a).

Our narrative approach is motivated by recent research in narratology that is highly influenced by cognitive science, as well as from the structuralist tradition defined by Vladimir Propp (1928) and others. Edward Branigan, one of the fathers of the cognitive science-motivated narratology, writes that narrative is a "perceptual activity that organizes data into a special pattern which represents and explains experience" (Branigan 1992). It is this understanding of narrative that we use to develop our ideal narrative function. Rather than trying to model a text out of Wikipedia, which possesses a certain set of traits (such as characters who have goals), our narrative algorithm shapes a disparate text corpus according to characteristics of a desired narrative *experience*.

It is from David Bordwell's work (2006) that we define two of the most important of these characteristics: unity and development. Hecht et al. (2007a) describe Minotour's use of unity as follows:

> "According to Bordwell (2006), a precondition for understanding a fiction film as narrative is the unity of the text as a formal system. He writes that, in the most formally unified films, 'every element present has a specific set of functions, similarities and differences are determinable, the form develops logically, and no element is superfluous. In turn, the film's overall unity

> gives our experience a sense of completeness and fulfillment' (Bordwell 2006). Thus, for a reader or viewer to understand the narrative as unified, they must perceive the individual elements as interrelating and nothing must seem 'out of place.'"

Bordwell (2006) defines development simply as a "progression moving from beginning to middle to end" (Hecht et. al 2007a).

It is important to note that simply including unity and development in a narrative will not magically turn it into a narrative. Rather, these are important narrative cues that will encourage the user to execute story construction activities (Bordwell 1987) and utilize narrative schema while reading the text. In other words, unity and development will help to elicit a narrative *experience* in the user.

2.4.1 Unity in the Ideal Narrative Function

Taken as a whole, Wikipedia is a disparate collection of facts (and some opinions) with no inter-article coherence. However, as noted above, unity is a critical narrative cue. As such, unity must be an important characteristic of the ideal narrative function, the function against which all possible narratives between two 3D articles are evaluated. Of course, the most obvious way that we provide unity is by linking 3D article a to 3D article b through a series of other articles. Before the user reads the narrative, these articles were likely perceived as distinct entities; afterwards, they are inherently connected. But this form of unity is weak and only informs our ideal narrative function in the most basic way.

More significantly, we provide unity by highlighting the themes in the user's space. This approach to unity is informed by our educational tourism focus. A critical element of geography education is the highlighting of themes embedded in the geography of a region. In fact, a key element of regionalization – a backbone of the United States National Geography Standards (National Geographic Research & Exploration 1994) – is being able to reduce the complexity of a diverse area through the construction of thematic regions.

In (Hecht et. al 2007), it is noted that,

> "While it might seem that incorporating a number of themes might disconnect the narrative, the presence of multiple themes can unify the narrative experience as a whole. Each theme serves to draw a conceptual link between two or more snippets. For example, in our sample narrative, issues about German freedom, unification, and community are directly referenced in the snippets about the 'Brandenburg Gate,' 'Hans Modrow,' 'Strausbourg,' and 'Humbolt University of Berlin.' The themes of German freedom, unity, and community thread together these snippets. On the other hand, the intersection of the academic institution with political and religious forces is present in the snippets from 'Strausbourg,' 'Martin Luther,' and 'Humbolt University of Berlin.'"

The need to quantify the thematic requirement is the reason we chose in-links as the defining variable in the ideal narrative function. In the context of Wikipedia, in-links are a highly accurate proxy for generality. When an author working on one entry links to another entry, that author is essentially saying "this entry is important to my entry". Articles that are more important to more entries are more general articles.

For instance, in the German Wikipedia, the article for "Poker" has far more in-links than the article for the "1990 World Series of Poker". We make the assumption - proven to be mostly true experimentally - that more general articles have more thematic content. Given the in-links/generality/theme relationship and the ideal narrative function's focus on high in-links articles, the reason for the appearance of themes in our ideal narrative function becomes clear. It is important to note that we include in the narrative function a maximum number of in-links that is less than the actual maximum number of in-links. We found through experiments that articles with extremely high in-link totals are too broad to carry much interesting thematic significance.

If we were just focused on theme, it would be ideal to find a set of articles with a high number of in-links (but not too high) and present these to the user. However, it is critical to the narrative foundation of our tours that the reader understands the snippets of the text as developing in a certain direction.

As was the case with unity, there is an obvious answer to the question of how our ideal narrative function incorporates development: each snippet is linked to the next snippet. However, just as with unity, the function also includes conceptually deeper models of development. We accomplish this by incorporating a small positive slope in the first two-thirds or so of the function. In this part of the function, each snippet becomes only slightly more general than the next. As such, the reader gains a sense of movement toward generality. She can then question and form hypotheses about the broader themes of her space and how they will connect her start to her

destination. This activity mirrors the experience of the traditional narrative reader. In the traditional narrative text, conflict builds to a climax, followed by a resolution. In the context of our ideal narrative function, this conflict is manifested in the question "What does this snippet from this thematic article have to do with the specific space in which I am moving"? As such, before the user is returned from wiki-space to a 3D article, the broadest theme of that space – the climax – is revealed. Then, the movement from the broadest theme back to the concluding 3D article – the resolution – provides a sense of completeness and closure to the experience. In the case of our sample narrative in figure 1, the climax occurs at the Martin Luther snippet. At this point, we expect that the user is maximally wondering what a highly broad article like "Martin Luther" has to do with the specific space in between the Brandenburg Gate and Unter den Linden. This curiosity is quickly satisfied in a mere two hops through the Wikipedia graph, as the user learns of the connection through Humboldt University.

Since Minotour's generated texts are designed to be delivered as the user travels from feature a to feature b, they are accompanied by the plotline the user is experiencing by moving through the space. The progression of this two-trajectory context in which the stories diverge at the beginning (at a) and rejoin at the end (at b), is one that the user is used to perceiving within a narrative, and thus aids development. Thousands of examples of this context exist in pop culture media. For instance, nearly all *Seinfeld* episodes begin with one narrative that splits into two (or more) at the beginning and intersects at the end ("The Limo", etc.)

2.5 Implementation

We have taken a basic client/server approach to our implementation, an approach that maximizes the individual flexibility of the client and the server. All of the narrative generation work takes place on the server, while all of the narrative delivery is done by the client. When a narrative is desired, the client (currently a Windows Mobile 5 application still in very early stages of development), sends the article id number of 3D articles a and b to the server, and the server returns the optimal narrative between these two features.

The server operates on information provided by a MySQL database of Wikipedia data parsed by WikAPIdia. The client side of this project is the lesser-developed side, and we are exploring delivery platforms other than handheld devices. That said, our current handheld software, developed in C# and the Windows Mobile Native C++ API, is effectively a mini-geographic information system (GIS) with GPS capabilities, customized for the various features and desired user experience of Minotour. Because no suitable open-source code could be located, it was necessary to write this software from scratch. A screen shot of a software-emulated version of the application can be found in figure 2d.

Figure 2d: A screenshot from a software emulation of our Minotour handheld client.

We have also implemented Minotour AJAX, a desktop online client to the server-based narrative generator (figure 2e) . The tool, built with Google Web Toolkit, offers a wide variety of adjustable parameters (from minimum and maximum snippet size to number of snippets to whether or not to include spatial articles in the body of the narrative) and allows users to select start and end features. While having certain different narrative experiential properties, is a useful tool for testing as well as for narrative theory researchers exploring Minotour.

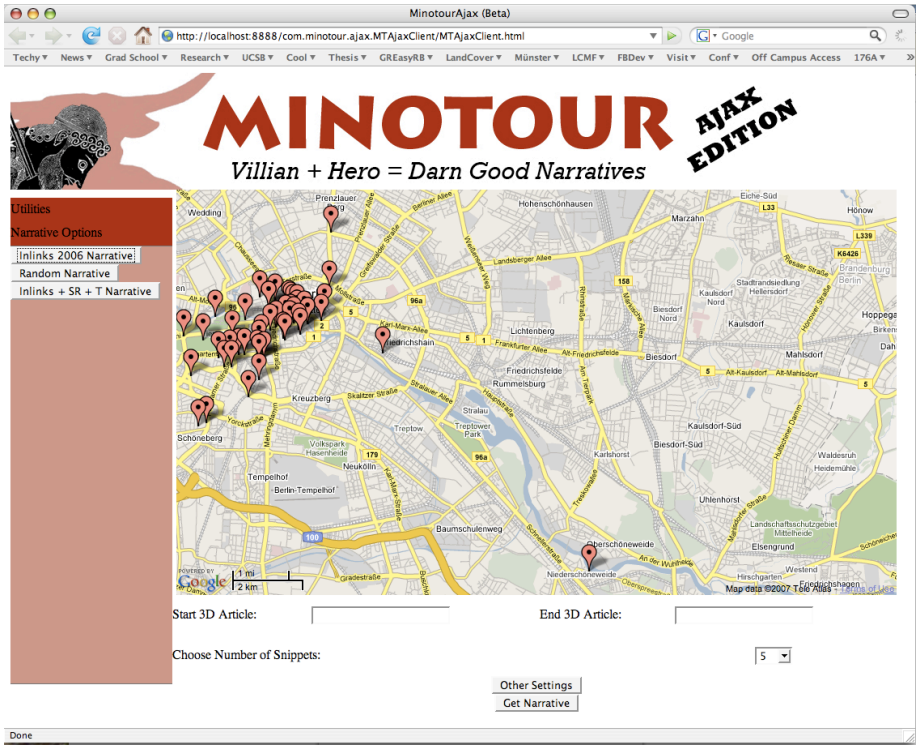


Figure 2e: A screenshot from Minotour AJAX, the online implementation of Minotour. It uses both Google Web Toolkit and Google Maps.

2.6 WikEar - An Audio / Magic Lens Extension of Minotour

WikEar (Schöning et al. 2007) is a marriage of Minotour's content with TIMMI's (Schöning et al. 2006) interface. It is the first application of Minotour and is targeted more directly for use in the tourism industry than is Minotour. Currently, tourists who want an educational experience during their vacations have two do-it-yourself options – paper guidebooks and customized, highly localized mobile device applications – both of which have severe content limitations and are not available for many locations. Writing, editing and post-production of content in these types of

tourism tools can be expensive and overwhelming. In addition, the content can quickly become out-of-date. WikEar is an attempt to wed the pervasive and easily-updated content of a mobile map application with the educational capabilities of a guidebook. In both of these goals, the narratives generated by Minotour are critical.

The premise of WikEar is quite simple: The user stands in front of a public city map and selects a spatial feature (such as a building or landmark) using her camera phone as a magic lens, as described in (Schöning et al. 2006). A guided audio, narrative-based tour between the location of the city map and the destination is then delivered to the user, with the intent that she will listen to the story as she travels to the destination. To track the mobile device relative to the map we use the magic lens tracking technology in (Rohs et al. 2007), which combines the high-resolution visual context of paper maps with the dynamic information capabilities of mobile technology. The guided tour comes in the form of a narrative that is automatically mined from Wikipedia by the Minotour system. In the future system, the output will render in audio form using text-to-speech (TTS) technology. However, for the demo implementation produced, I recorded twenty sample narratives from Minotour.

The twenty narratives were originally generated from the German Wikipedia and then were translated into English by Johannes Schöning, a co-author on the WikEar paper (Schöning et. al 2007). This was done because the conference at which the demo was to be shown, UbiComp 2007, was located in Innsbruck, Austria, and we wished to present users of the device with an in-context set of examples.

2.7 *Evaluation*

While we have not yet completed a planned user study on the effectiveness of Minotour in eliciting narrative schema in its users, we encouraged UbiComp 2007 attendees who demoed WikEar to fill out an online questionnaire about their experience. This evaluation had two key questions related to Minotour borrowed from the draft version of our evaluation:

1) How cohesive were the stories? Did all the pieces seem to make sense together?
2) Was there a sense of development from beginning to middle to end? Did you feel that the story was progressing to a specific point?

Obviously, question one was an attempt gauge how unified the users perceived the narratives to be. Question two did the same for progression. Both questions were to be answered on a six-point scale. A "1" on question one indicated that the stories were "unified" and a "6" indicated that the stories were "completely disjoint". Similarly, a "1" on question two represented that the the user found that the "story developed as [they] read" and that they often thought about "what might happen next". A "6" on the same question signaled that the "felt lost" and that the story had no clear direction.

Out of twenty-one responses to the online survey, twenty users sent in valid data. Because the standard deviation on both questions was so high, very few *conclusive results* can be drawn. The mean of question one was 2.38, but because the

standard deviation was 1.47, the 95 percent margin of error covered nearly the entire range of possible values. The same occurred with question two, which had a mean of 2.76 and a standard deviation of 1.44. However, the fact that a large majority of people report positive results of "1s" and "2s" is very encouraging. Figures 2f and 2g display the full histogram of results.

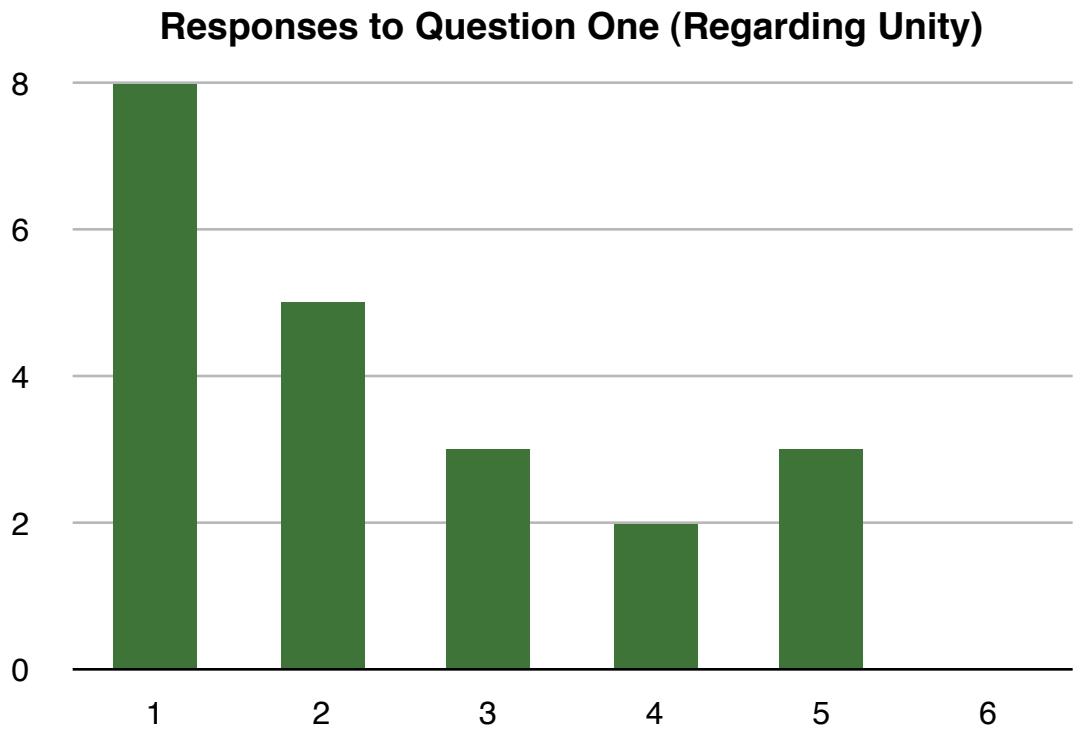


Figure 2f: Responses from our informal user study about the unity of Minotour narratives. A "1" indicates "unified" and a "6" indicates "very disjoint".

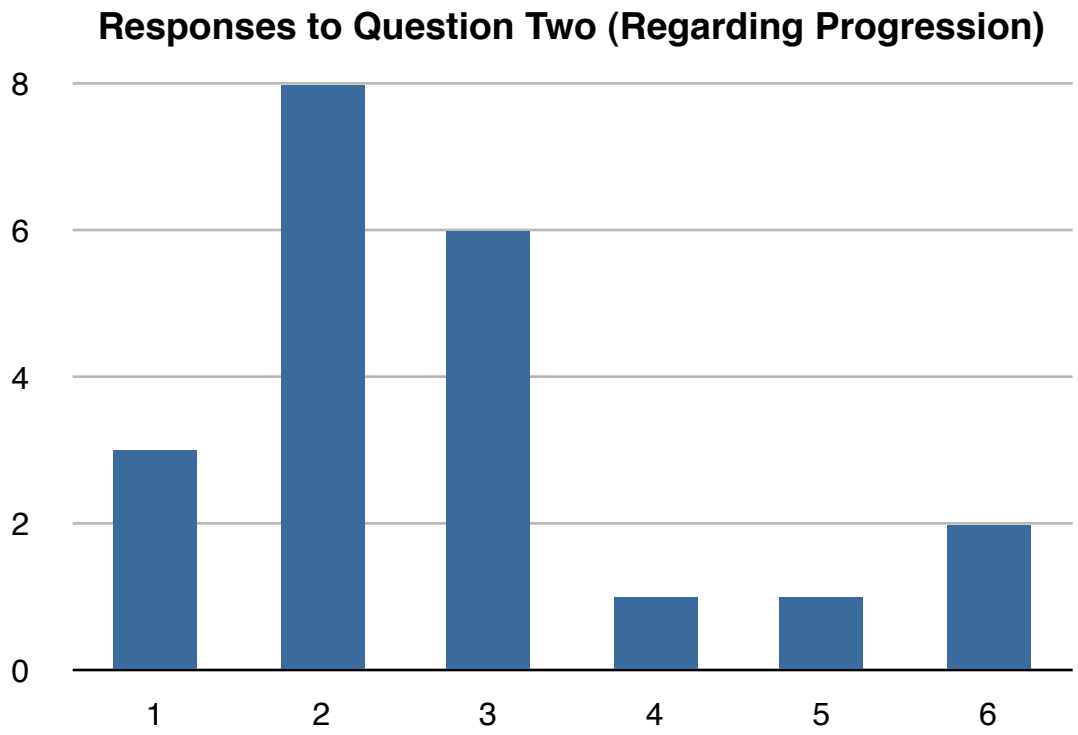


Figure 2f: Responses from our informal user study about the progression of Minotour narratives. A "1" indicates a developing story and a "6" indicates feeling lost.

2.8 Future Work

In this section, we have presented a novel, narrative theory-informed approach to data mining from a Wikipedia corpus within the context of educational tour generation. While we have completed a proof-of-concept system, we have many more research avenues to explore.

First and foremost, we intend to rigorously investigate both theoretical and artistic methods for increasing the narrative pleasure of our generated narratives. One idea currently in the works is eliminating spatially-referenced articles (in addition to

temporally-referenced articles) from the body of the narratives. Spatially-referenced articles play a critical role as the start and the end points of our narratives. However, we have found that when they have snippets that appear in the body, these snippets suffer from the same semantic weakness problems as purely-temporally referenced articles. For instance, in figure 2a, the implicit "lies on" relationship between Humboldt University and Unter den Linden is a rather uninteresting and non-educational one, and one that would be obvious to any user who looked at the map on their mobile device client. Other possible avenues for increased narrative pleasure include providing theme-based tours (by utilizing Wikipedia's category graph) and using Wikipedia's link structure to include more articles in the narratives that are closely tied to the space in which the user is traveling.

Secondly, we are exploring different narrative theoretical constructs with which to examine and improve our generated narratives. We aim to develop further research in the area of cognitive science-based narrative theory with the goal of proposing better strategies for the interdisciplinary adoption of contemporary narratology.

We are also excited about the possibilities for client-side development. Raw text presentation is not the best delivery format for our narratives. We are exploring many possibilities for ways to utilize the presentation as a means to aid narrative interpretation, either through the increase of unity and development, or via other theoretical constructs. One idea we have explored is found in WikEar (Schöning et al. 2007). We are also looking into turning the faceless narrator of our stories into a

more explicit character, perhaps even similar to that in Persson et al. (2003) or Isbister and Doyle (2003). We have many ideas for using the properties of Wikipedia to enact Persson's suggestion to closely integrate character behavior and presented content. This might include placing greater emphasis on the real life Wikipedia users who input content, as well as considering article conflict statistics.

Following Persson et. al (2003), we are taking steps towards designing an appropriate evaluation for Minotour to extend our informal study. We will determine the average user's reliance on narrative schema for the comprehension of a random series of linked snippets (our baseline) as opposed to Minotour generated narratives.

Finally, we also plan on implementing spatial feature contribution functionality. Ideally, users should be able to enter information about a spatial feature on which Wikipedia has no information and, immediately afterwards, receive a tour with that feature as the start or end destination. This would add a whole new level of interactivity to our current system.

Minotour is currently heavily customized towards its educational tour narrative generation test platform. However, the ideas behind the implementation can be utilized in other areas and with different goals. Most apparently, the concept can be applied to generating narratives between any two Wikipedia articles, spatial or not. In initial tests, the authors learned a lot from narratives generated between biographies and even between two mathematical concepts. More broadly, we are considering how this idea could be employed on the Internet as a whole. While we

have not explored the narratology implications of such non-spatial and/or non-Wikipedia applications, these are vital further research directions.

Chapter 3 – WikEye: Using Magic Lenses to Explore Georeferenced Wikipedia Content

3.0 Abstract

Traditional paper-based maps are still superior in several ways to their digital counterparts used on mobile devices. Namely, paper-based maps provide high-resolution, large- scale information with zero power consumption. Digital maps offer personalized and dynamic information, but suffer from small outer scales and low resolutions. In this paper, we present WikEye, an interdisciplinary project that demonstrates proof-of-concepts of three novel research ideas related to this topic: multi-dimensional interaction with a magic lens device, spatiotemporal Wikipedia data mining, and magic lens markerless tracking. Through its integration of the advantages of static and dynamic maps and its employment of Wikipedia as its primary knowledge repository, we envision WikEye being used to address mobile technology issues in the tourism and education fields. In particular, we focus on the need for effective tourism-related mobile technologies and the demand from educators for new methods that students can use to interact with information. Though the full paper on WikEye (Hecht et al. 2007b) elucidates all three research ideas, this thesis chapter will focus on my main contribution to the project: the spatiotemporal Wikipedia data mining.

3.1 Introduction

Tourism provides over six percent of the world's gross domestic product

(UNWTO 2003). However, despite the fact that mobility is at the heart of tourist activity (Hall 2005), mobile technologies have yet to infiltrate the tourism industry (Brown 2003). Similarly, educators are always seeking out new technologies for student information interaction, such as in (Leichtenstern and André 2006) . Applications like RFID maps (Reilly et al. 2006), Timmi (Timmi is Mobile Map Interaction) (Schöning et al. 2006a, Schöning et al. 2006b) and Minotour (Hecht et al. 2007a) are members of a new generation of systems aimed at making mobile technology more appealing to tourists, and Minotour is designed specifically for tourism-based education. Timmi combines the advantages of large scale and high-resolution maps with the interactivity and up-to-date nature of dynamically queried geoobjects. Through the integration of the premise of Timmi and its new markerless tracking interface with the geography education-oriented Wikipedia technology and research of Minotour, we can provide users of mobile devices with a novel approach to space and place understanding through dynamic display empowered static maps.

Like Semapedia[5] and similar projects, WikEye takes Wikipedia off the computer and moves it into the real world. However, WikEye does so in an entirely novel fashion: by placing Wikipedia data into interactive spatial, temporal, and semantic augmented reality contexts that are referenced to static maps. In doing so, we meet the call for greater interaction between paper maps and guidebooks (in this case, Wikipedia) in (Brown 2003), providing a prototype mobile technology appealing not only to the tourist, but to the educator as well.

---

5 http://www.semapedia.org/

Because of its Wikipedia article density and its history, a city map of Berlin provides an excellent test case for WikEye. Maps of Berlin alone cannot satisfactorily educate the user in the rich history of Berlin, and using a map and another information source at the same time can be a cumbersome experience (Brown 2003). Ideally, visitors to Berlin or students of the city would be able to easily view information contained in a book or travel literature through the context of a large-scale, high-resolution paper map. WikEye provides this exact functionality in an easy-to-use handheld application powered by a novel map-device interaction scheme and informed by the vast quantity of free information in Wikipedia.

WikEye has three main components, each of which comprises an innovative area of research. This thesis chapter only covers detailed information on the Wikipedia portion of the research, which is discussed in section two. In section three, the reader will find a brief summary of the other two main research components. Implementation details are summarized in section four. Finally, a conclusion and WikEye's relationship to later research, especially GeoSR, is discussed in the closing section.

3.2 Wikipedia Data Mining for WikEye

As noted in the introductory chapter of this thesis, Wikipedia has been extensively studied as a phenomenon, but its utility as a vast repository of free world knowledge is little understood (Gabrilovich and Markovitch 2006). In this project, we mine Wikipedia for three key types information: spatial data, temporal data, and

semantic connections between places (see figure 3a). Without such a massive and freely available data corpus, the utility of WikEye would be categorically diminished, and WikEye would require significant and impractical specialized content development.

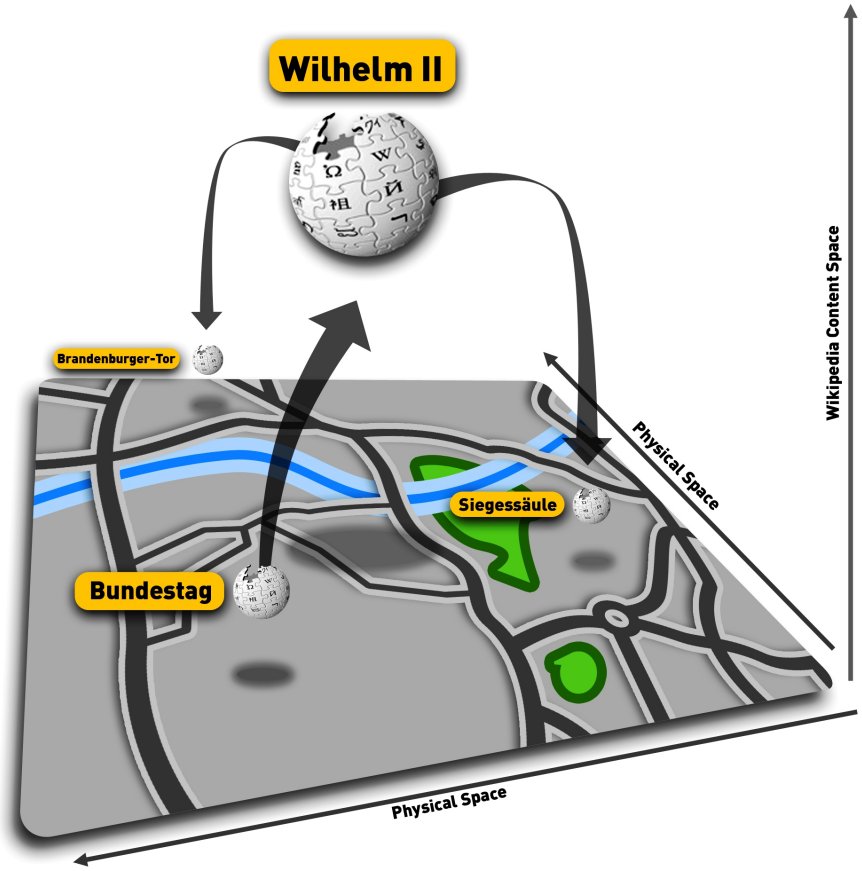


Figure 3a: Spatial feature-edge-feature relationships in Wikipedia.
Graphic by Johannes Schöning (Hecht et al. 2007b)

Wikipedia spatial data extraction is done as is described in chapter one. The goal of our Wikipedia temporal information extraction procedure is to discover the start and end years of spatial features via the temporal expressions in the features'

corresponding Wikipedia articles. Critically, key historical eras in each article must also be identified. The mining of natural text for temporal information is a well-understood problem (Mani 2004), but it has proven difficult to solve with high levels of accuracy (see introductory chapter for more details). As is noted above, however, due to the certain properties of Wikipedia, we have found that we can gather temporal information with both sufficient quantity and accuracy by using relatively simple temporal information extraction procedures, such as those described in (McKay and Cunningham 2000). In addition, temporal accuracy is improved by the pre-disambiguated temporal expressions in Wikipedia. These disambiguations take the form of wiki links to the large number of articles on specific years, such as the "1983" article, for example.

Once the temporal information is extracted, the application must then identify the start and end years of each spatial feature, as well as features' key historical periods and/or eras. This is done via the analysis of article temporal reference profiles with pattern recognition techniques and other statistical methods. Temporal reference profiles are essentially histograms that describe the number of references to each year on a timeline. Figure 3b shows the profile of the article on Berlin.

The fourth dimension of interaction for our application is spatially-referenced semantic connections. Mining Wikipedia for semantic connections between places is a simpler problem than temporal and spatial information extraction. Of course, easily identified (although not easily interpreted [Völkel 2006]), semantic connections are a key benefit of the Wiki concept. Due to the restraints inherent to our novel interaction

scheme (see below), it was determined that only spatial feature-edge-node-edge-spatial feature relationships (e.g. Bundestag - Wilhelm II - Brandenburg Gate) should be considered (see figure 3a). In other words, only relationships between two spatial features through a single intermediary non-spatial Wikipedia article are presented to the user. Spatial feature-edge-spatial feature relationships (Berlin - Bundestag) are not included, as they have been found to be overloaded with the spatial relationships that are already evident to the user via the map display. Delivery of the "story line" between the features and through the intermediary article is similar to that described in the Minotour chapter.

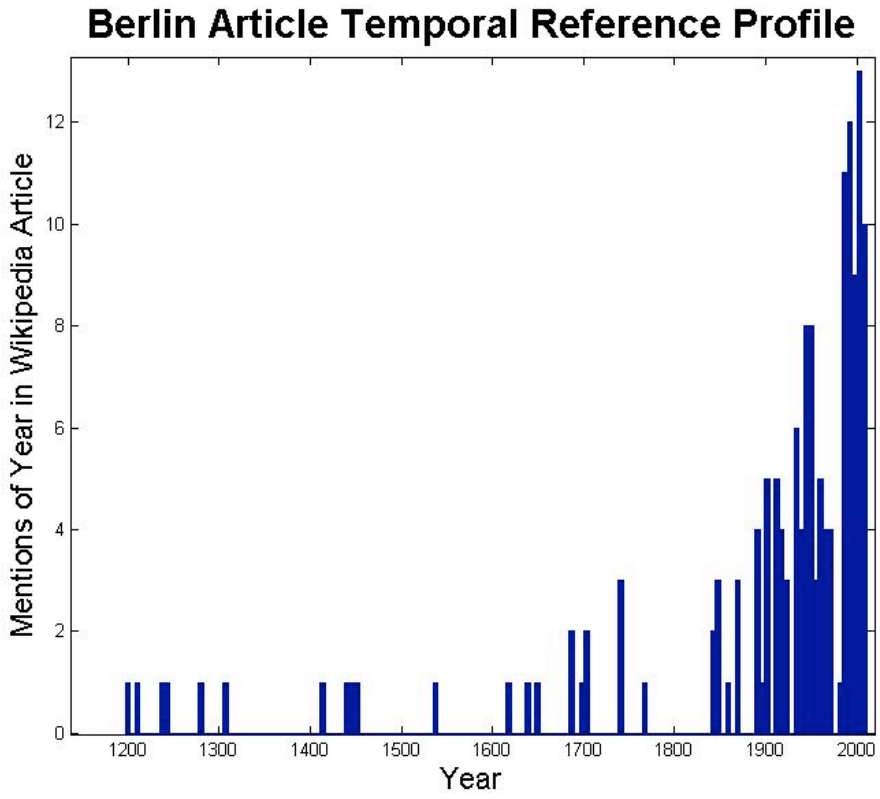


Figure 3b: The Berlin article temporal reference profile (from the German Wikipedia)

3.3 Summary of WikEye Interaction Research

Note: Although I was a strong participant in this portion of the research, it was not led by me. As such, this small section is provided simply for context.

By using a mobile device and making explicit the spatial, temporal, and relational structures implicit in Wikipedia, we free the encyclopedia's data not only from the constraints of the desktop but also from its rigid interaction medium. The goal of the interaction portion of the research is to define an interaction paradigm for these newly available structures.

As is noted in the introduction, the overall interaction approach is defined by our choice of the cell phone-based magic lens technology. This approach allows the user to utilize her or his camera-enabled cell phone to provide a "magic lens" view of a printed map. It is in this "magic lens" context that WikEye must provide an interface to the aforementioned Wikipedia structures. Specifically, we enable the browsing of Wikipedia via for entirely new interaction tasks:

1) Space-time exploration
2) Space-time selection
3) Space-Wikipedia relationship exploration
4) Space-Wikipedia relationship selection

Each of these tasks have been constructed using the interaction primitives defined by Rohs and Zweifel (2005). For instance, by *sweeping* the mobile camera device over the map, the user can explore Wikipedia articles on the map (space

exploration). By *rotating* the device – evoking a clock metaphor – the user can simultaneously wander the chronological eras relevant to the articles within the current spatial extent (temporal exploration). A more specialized task is the space-time selection of an object or area with *point* and *shoot* (*keystroke*) in order to view Wikipedia text data from a spatiotemporal region. Again, by rotating the mobile camera device the user chooses a time interval. By clicking (*keystroke*) with the cross-hair on a georeferenced Wikipedia object the user gets a Wikipedia paragraph, or "snippet", from the object for the selected time period.

For example, while exploring a map of Berlin with WikEye, a user may want to know more about the Reichtstag, a spatial article tuple extracted in the Wikipedia data mining process. If the user wishes to know more about the Reichtstag during the beginning of the Second World War, the user selects this period by rotating the device to the correct time period and clicking on the Reichstag's icon. The user would then receive a Wikipedia snippet about the Reichstag during that period.

A second basic interaction paradigm is that of space-Wikipedia relationship exploration. The user interacts with space-Wikipedia relationships by sweeping the mobile device over the map to view arrows drawn on the screen between related georeferenced Wikipedia articles (figure 3c). These arrows have labels that attempt to succinctly describe the relationships they represent, relationships that are limited to that of feature-edge-node-edge-feature connections. Space-Wikipedia relationship selection is a derivative of space- Wikipedia relationship exploration. By selecting a georeferenced Wikipedia article with a click, the user can explore feature-edge-node-

edge-feature relationships beginning at the selected feature. By selecting a second object, a snippet- based “story line” between the two articles is displayed on the mobile device, as is done more elaborately in Minotour.

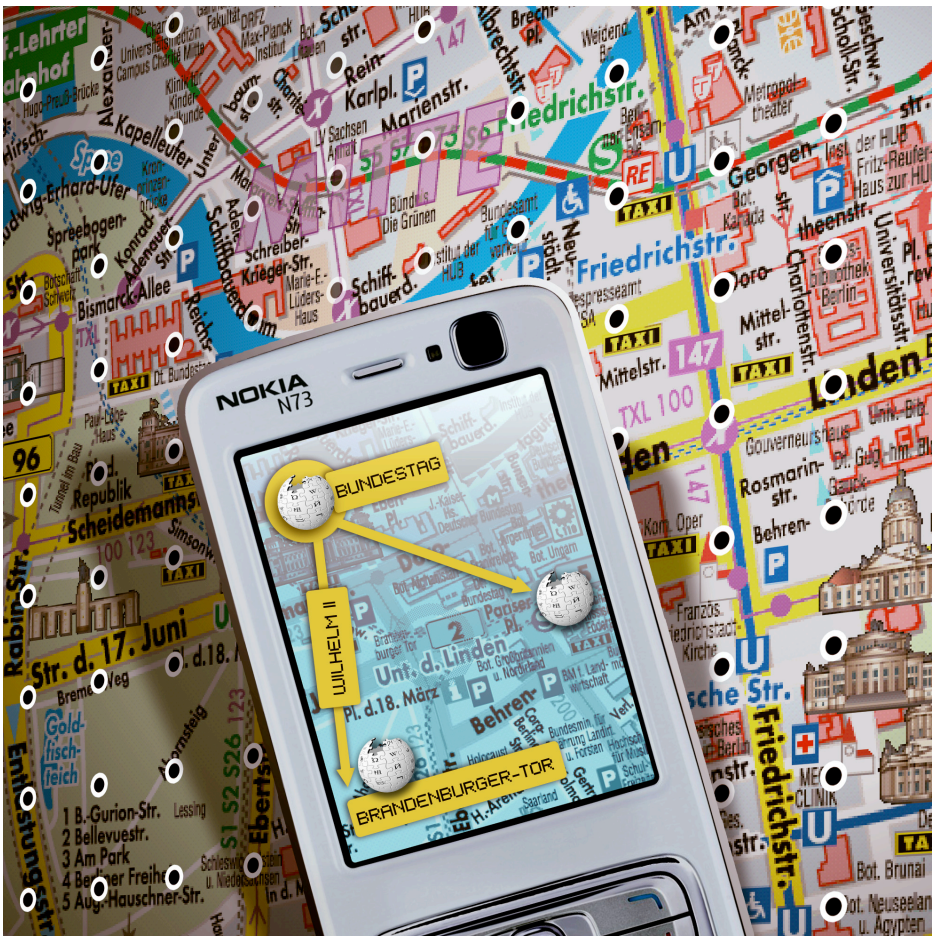


Figure 3c: Application in use (mock-up)

3.4 Implementation

All Wikipedia processing takes place within WikAPIdia, which also serves as an explicit API back-end to all the Wikipedia structures mentioned above. Our “client” device is of course a camera-enabled mobile phone, specifically a Nokia N80.

In our demo implementation, self-contained Wikipedia data sets are preloaded onto the phone, but in a real-world version these data sets would be prepared dynamically and downloaded via a network from a server running an iteration of WikAPIdia. More implementation details can be found in the full paper (Hecht et al. 2007b).

3.4 Conclusion

While we have successful demonstrated a proof-of-concept tool that merges the guidebook and the map in a manner that would benefit the tourism and education markets, many questions still exist. In the data mining portion of this project, these questions include determining the optimal temporal temporal information extraction methodologies and developing better event detection algorithms, possibly following (Smith 2002).

In addition to its valuable contributions to the tourism and education technology experience, WikEye has also been vital in its role as a first step toward the research present in the next chapter of this thesis, research whose contribution could have effects well beyond the tourism and education contexts. In fact, GeoSR, the subject of the next chapter, can in a way be seen as the logical generalization of the data mining-related ideas developed in WikeEye. Namely, GeoSR also restructures Wikipedia data "along spatial, temporal, and semantic dimensions" (Hecht et al. 2007b, p. 4) but it does so in such a way that provides a much more generalized and powerful data exploration environment.

Chapter 4: GeoSR - Geographically Explore Semantic relations in World Knowledge

4.0 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.

4.1 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 through 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 doing so.

GeoSR uses Wikipedia as its knowledge repository. 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 the 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.

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 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 measure 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 data sets) 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").

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. 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 4.1a 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 “Tuxpan (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 4.2b. 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 five.

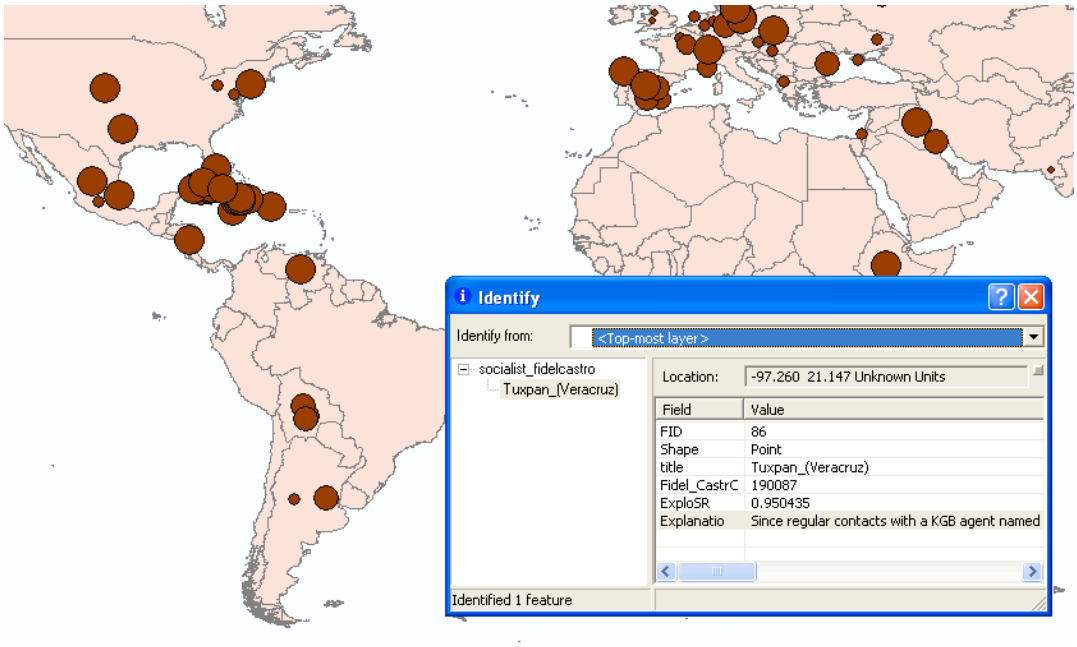


Figure 4.1a: A visualization of GeoSR data in which Fidel Castro was the input entity. Large dots represent the most related locations to Fidel Castro and smaller dots represent less important locations (within the top 100 locations).

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 4.1b: An expansion of the content of the explanation field seen in figure 1.

Section two of this chapter describes the advantages of using the Wikipedia Article Graph (WAG) over other structures in the encyclopedia in this context are discussed in section three. Section three covers ExploSR in detail, highlighting its strengths and weaknesses. In section four, 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.

4.2 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 a spatial feature and non-spatial entity. The rest of this section is dedicated to explaining these two advantages in detail.

4.2.1 The Wikipedia Snippet - Paragraph Independence Facilitates Data Exploration

Note: Although some of this information is repeated from chapter one, it is included as it is absolutely critical to the understanding of this chapter.

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 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 means that the meaning of a link is almost always contained within the snippet that contains the link (see figure 2). Additionally, this property ensures that snippets can be safely rearranged or presented independently without severely reducing their information content. We have found that the only context necessary for fully understanding 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 the 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 highly exogenous to the relatedness measure.

4.2.2 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 current lexical resource focus of most semantic relatedness researchers, 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 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 *Silicon Valley* 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 there 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.

4.3 ExploSR: Using the WAG for Semantic Relatedness

4.3.1 Microstrucutre 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 an 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 relevance measure, or weight, to each edge in the WAG. To do so, it uses the following general formulas:

If $|OL_A| > C$,

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

Else,

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

And if $|OL_B| > C$,

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

Else,

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

with the final ExploSR value being,

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

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*. $|OL_{A \to B}|$ and $|OL_{B \to A}|$ 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[6].

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

[6] 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.

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*.

4.3.2 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 to the two entities involved 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.

4.3.3 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. 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 along 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 4.1a), along with the attribute data described in the next section.

4.4 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.

4.4.1 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 4.1a and 4.4a. 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 4.1b).

We have implemented this system by exporting the output of GeoSR into a shapefile and loading this data into ArcGIS[7]. The shapefile contains three columns in its attribute table: name of the spatial entity, its ExploSR value, and its snippet path. 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).

7 http://www.esri.com

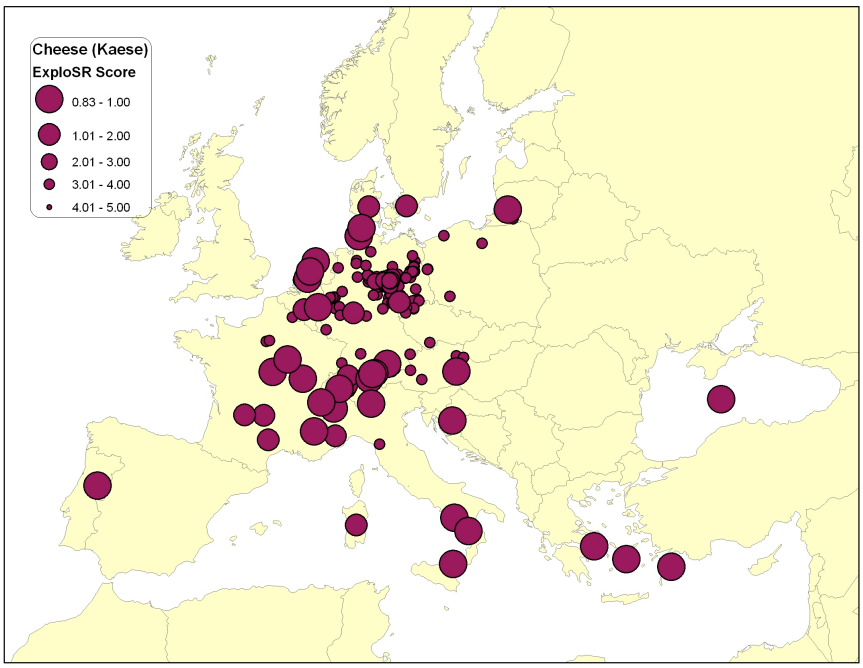


Figure 4.4a: A visualization of the output resulting from inputting the article *Käse* (German for *Cheese*) into the GeoSR system operating on the German Wikipedia. Spatial stemming was turned on, and missing links were not included. The top 200 locations were output, but not all are located in the region depicted above.

4.4.2 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 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.

4.4.3 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 a spatial article as the input, have another means of exploring the degree to which this guideline holds true.

4.4.4 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.

4.4.5 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.

4.5 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, the temporal reference system 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 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.

Chapter 5: Conclusion

Although each chapter of this thesis has its own conclusion with a description of individual project contributions and a summary of future work, it is important to examine the meta-effects of this research as well as take a broader look at the research that lies ahead. Both of these can be elucidated easily: The main contribution of this work is that it represents the first-ever research on Wikipedia that has been informed by a spatiotemporal viewpoint. Future work should include the study of Wikipedia by all sub-fields of Geography, from the geoinformatics perspective taken here to the more critical humanities areas of the discipline. An incredible number of geographic questions remain about this fascinating phenomena, as well as others like it that may crop up in the future. The rest of this section will describe just a small portion of these questions.

During the course of this research, I encountered many geoinformatics-oriented questions that I was unable to answer due to the desire to keep this thesis focused on a small number of projects. One such question involves investing the degree to which articles about spatial phenomenon are written by people near to that spatial phenomenon. In other words, this research would test the degree to which the first law of geography holds for Wikipedia contributions. A colleague and I also plan to examine whether spatial articles written by Wikipedians located close by to

(identified via IP address) to these articles tend to be more trustworthy and/or accurate. A similar question would involve vandalism and would ask to what degree vandalism on spatial articles is influenced by the distance of the miscreant to the vandalized articles. In other words, is local or distant vandalism significantly more common than the other?

The question of why certain Wikipedias have such high Wikipedia articles / number of language speakers ratios and other have such low ones is definitely in need of study by a quantitative human geographer. For instance, the Portuguese Wikipedia is significantly larger than both the Spanish Wikipedia and the Arabic Wikipedia, and the Dutch Wikipedia is significantly larger than that of Portuguese! Also, the same group of geographers should look at why certain areas of the world are densely covered with spatial articles by the English and German Wikipedias, but others are not. More qualitative sub-fields of human geography may also have interesting perspectives on these topics.

Finally, a critical viewpoint is needed to analyze the space and place issues of Wikipedia. Primarily, what does it mean to have such broad, anyone-goes collaboration on such an important public description of spatial features. Are certain groups getting more exposure than others? Are the groups marginalized by the digital divide becoming invisible in Wikipediaplace? These are all issues in dire need of further study.

Acknowledgements

The work presented in this thesis could not have been possible with the assistance and encouragement of a large number of friends, colleagues, family, and professors. Many of these names are listed below:

My committee (Dr. Keith Clarke, Dr. Martin Raubal, Dr. Tobias Höllerer), Dr. David Lanegran, Dr. Laura Smith, Dr. Antonio Krüger, Dr. Michael Rohs, Johannes Schöning, Nicole Starosielski, Shaun Stehly, Drew Dara-Abrams, Kirk Goldsberry, Julie Dillemuth, John Roberts, Clare Salustro, Emily Moxley, Meri Marsh, Ellen Hecht, Barry Hecht, Jory Hecht, Garin Hecht, Herb Hecht, Esther Hecht, Sam Nalle, Mara Brady, Cara Harwood

References

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

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

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

Bednarz, S. W., Bettis, N. C., Boehm, R. G., De Souza, A. R., Downs, R. M., Marran, J. F., et al. (1994). National Geography Standards. Washington, D.C.: National Geographic Research & Exploration.

Bordwell, D. (1987). Narration and the Fiction Film. London: Routledge.

Bordwell, D. & Thompson, K. (2006). Film Art: An Introduction (8th ed.). Boston, MA: McGraw Hill.

Branigan, E. (1992). Narrative Comprehension and Film. London: Routledge.

Brown, B. & Chalmers, M. (2003). Tourism and mobile technology. Proceedings of the Eighth European Conference on Computer Supported Cooperative Work, Helsinki, Finland.

Budanitsky, A. & 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., & Millozii, S. (2006). Temporal Analysis of the Wikigraph. Proceedings of Web Intelligence, Hong Kong.

Cherones, T. (1992). The Limo On Seinfeld.

Colbert, S. (2006). The Steven Colbert Show - July 31, 2006 New York, NY: Comedy Central.

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

Forte, A. & Bruckman, A. (2005). Why Do People Write for Wikipedia? Incentives to Contribute to Open-Content Publishing. GROUP 05.

Gabrilovich, E. (2007). Wikipedia Preprocessor (WikiPrep). http://www.cs.technion.ac.il/~gabr/resources/code/wikiprep/

Gabrilovich, E., Markovitch, Shaul. (2006). Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. Proceedings of the 21st National Conference on Artificial Intelligence (AIII-06), 1301-1306.

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

Gretzel, U. & Fesenmaier, D. R. (2002). Building Narrative Logic into Tourism Information Systems. IEEE Intelligent Systems, 59-61.

Hall, C. M. (2005). Reconsidering the Geography of Tourism and Contemporary Mobility. Geographical Research, 43(2), 125-139.
Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A Formal Basis for the Heuristic Determination of Paths in Graphs. IEEE Transactions on Systems Science and Cybernetics, SSC-4(2), 100-107.
Hecht, B. (2005). Mapalester: Powerful, Easy-to-Use GIS Software Under Development. Macalester College, Saint Paul, MN.
Hecht, B., Rohs, M., Schöning, J., & Krüger, A. (2007). 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., & Dara-Abrams, D. (2007). Generating Educational Tourism Narratives from Wikipedia. Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium on Intelligent Narrative Technologies, Arlington, VA.
Isbister, K. & Doyle, P. (2003). Web Guide Agents: Narrative Context with Character. In M. Mateas & P. Sengers (Eds.), Narrative Intelligence. (pp. 229-243). Philadelphia, PA: John Benjamins Publishing Company.
Kim, J.-W., Kim, C.-S., Gautam, A., & Lee, Y. (2005). Location-based Tour Guide System Using Mobile GIS and Web Crawling. Web and Wireless Geographical Information Systems. (pp. 51-63). Springer Berlin / Heidelberg.
Kühn, S. (2007). Wikipedia:WikiProject Geographical coordinates. http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung
Kuhn, W. (2003). Semantic reference systems. International Journal of Geographical Information Science, 17(5), 405-409.
Kuhn, W. & 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.
Lanegran, D. (2005). Discussion on question 'What Makes a Good Tour?'.
Leichtenstern, K. & André, E. (2006). Social mobile interaction using tangible user interfaces and mobile phones. Gesellschaft für Informatik, Mobile and Embedded Interactive Systems (MEIS) Workshop.
Liu, J. & Birnbaum, L. (2007). Measuring Semantic Similarity between Named Entities by Searching the Web Directory. 2007 IEEE/WIC/ACM International Conference on Web Intelligence (WI'07).
Mani, I. (2004). Recent Developments in Temporal Information Extraction. RANLP '03, Amsterdam.
Mani, I., Wilson, George (2000). Robust Temporal Processing of News. 38th Annual Meeting on Association for Computational Linguistics, Hong Kong.

Mateas, M. & Sengers, P. (2003). Narrative Intelligence (Introduction). In M. Mateas & P. Sengers (Eds.), Narrative Intelligence. (pp. 1-25). Philadelphia, PA: John Benjamins Publishing Company.

Mateas, M. & Sengers, P. (1999). Introduction to the Narrative Intelligence Symposium. Fall AAAI Symposium on Narrative Intelligence.

McKay, D., Cunningham, Sally Jo. (2000). Mining dates from historical documents. University of Waikato Department of Computer Science Technical Report.

Mehegan, D. (2006). Many contributors, common cause: Wikipedia volunteers share conviction of doing good for society. *The Boston Globe*.

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

Morris, J. & Hirst, G. (2004). 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).

Mott, B., Callaway, C., Zettlemoyer, L., Lee, S., & Lester, J. (1999). Towards Narrative-Centered Learning Environments. Proceedings of the AAAI Fall Symposium on Narrative Intelligence, Cape Cod, MA.

Persson, P., Höök, K., & Sjölinder, M. (2003). Agneta & Frida: Merging Web and Narrative. In M. Mateas & P. Sengers (Eds.), Narrative Intelligence. (pp. 245-258). Philadelphia, PA: John Benjamins Publishing Company.

Propp, V. J. (1928). Morphology of the Folktale. Austin, TX: University of Texas Press (1968).

Rainie, L. & Tancer, B. (2007). 36% of online American adults consult Wikipedia; It is particularly popular with the well-educated and college-age students. Pew Interent & American Life Project.

Reilly, D. F., Rodgers, M. E., Argue, R., Nunes, M., & Inkpen, K. (2006). Marked-up maps: combining paper maps and electronic information resources. Personal and Ubiquitous Computing, 215-226.

Riehle, D. (2006). How and why Wikipedia works: an interview with Angela Beesley, Elisabeth Bauer, and Kizu Naoko. WikiSym, Odense, Denmark.

Rohs, M., Schöning, J., Krüger, A., & Hecht, B. (2007). Towards Realtime Markerless Tracking of Magic Lenses on Paper Maps. Pervasive 2007 (Adjunct Proceedings), Toronto, Canada.

Rohs, M. & Zweifel, P. (2005). A conceptual framework for camera phone-based interaction techniques. Third International Conference on Pervasive Computing (PERVASIVE 2005), Munich, Germany.

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

Schöning, J., Heuer, J. T., Müller, H. J., & Krüger, A. (2006). The marauder lens. Fourth International Conference on GIScience (GIScience 2006) Extended Abstracts, Münster, Germany.
Schöning, J., Krüger, A., & Müller, H. J. (2006). Interaction of mobile camera devices with physical maps. Fourth International Conference on Pervasive Computing (PERVASIVE 2006).
Smith, D. A. (2002). Detecting and Browsing Events in Unstructured Text. ACM Conference on Research and Development in Information Retrieval, Tampere, Finland.
Strube, M. & 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.
United Nations World Tourism Organization. (2003). Tourism and the world economy. http://www.unwto.org/facts/eng/economy.htm
VanderWal, T. (2004). You Down with Folksonomy. http://www.vanderwal.net/random/entrysel.php?blog=1529
Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., & Studer, R. (2006). Semantic Wikipedia. International World Wide Web Conference, Edinburgh, Scotland.
Voß, J. (2005). Measuring Wikipedia. 10th International Conference of the International Society for Scientometrics and Informetrics.
Voß, J. (2006). Collaborative thesaurus tagging the Wikipedia way. Wikimedia Deutschland e.V.
Wells, C. G. (1986). The Meaning Makers: Children Learning Language and Using Language to Learn. Portsmouth, NH: Heinemann.
Wikimedia Foundation. (2007). Table of Wikimedia Projects by Size - Meta. http:// meta.wikimedia.org/wiki/Table_of_Wikimedia_Projects_by_Size
Wikipedia. (2007). Wikipedia. Wikipedia, The Free Encyclopedia http://en.wikipedia.org/wiki/Wikipedia
Zachte, E. (2007). Wikipedia Statiistics. http://stats.wikimedia.org/EN/TablesWikipediaEN.htm
Zesch, T. & Gurevych, I. (2007). Analysis of the Wikipedia Category Graph for NLP Applications. TextGraphs-2 Workshop (NAACL-HLT), Rochester, New York.
Zesch, T., Gurevych, I., & Mühläuser, M. (2007b). Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets. NAACL-HLT.
Zesch, T., Gurevych, I., & Mühlhäuser, M. (2007a). Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. Tuebingen, Germany.
Zlatic, V., Bozlicevic, M., Stefancic, H., & Domazet, M. (2006). Wikipedias: Collaborative web-based encyclopedias as complex networks. Physics Review E, 74(016115), 1-9.