

MEASURING RESEARCH DIRECTIONS IN GEOGRAPHIC INFORMATION SCIENCE USING LATENT SEMANTIC ANALYSIS

Submitted to the University Consortium for Geographic Information Science (UCGIS) summer assembly in Minneapolis, Minnesota, June 23-24, 2008.

David A. Parr
Department of Geography, Texas State University-San Marcos
601 University Drive, San Marcos TX 78666
Phone: 512-796-5715, Email: daveparr@txstate.edu

Abstract

This paper explores the semantic relationships between 770 Geographic Information Science research articles published between 1997 and 2007 in five different academic journals. Using keywords derived from the research priorities published by the University Consortium of Geographic Information Science (UCGIS), we extract closely related word stems to the original research priorities to provide a deeper, contextual search within the articles. For each research priority, closely correlated keywords are counted in each article, and a latent semantic analysis provides the correlated network relationship between UCGIS's research priorities, the year-by-year level of research in each subfield, and a distance-decay of similarity vs. distance. The ten most highly correlated locations of research for each UCGIS research priority are also found.

1 Introduction

1.1 UCGIS Research Priorities

In 1996, the University Consortium for Geographic Information Science (UCGIS) published their first set of white papers on research priorities for geographic information science (UCGIS, 1996). Since then, updates were published in 1998, 1999, 2000, 2002, and 2006. The consortium is a collaborative group of academic institutions, governmental organizations, and commercial GIS developers. In addition to recommending policy and legislation and setting goals for GIS education, the UCGIS advocates the advances in GIS research that are most important to its members. Listed in table 1 are the first set of published research priorities (UCGIS, 1996).

By updating their research priorities biennially, the UCGIS is providing the guideposts of future GIS research. Twelve years after their first publication, we can look back at the trends in the academic publications to see how the priorities define the subfields, advance the research, and pave the way for new advances in GIS. Now is the appropriate time to ask, 'has the academy responded to the challenges posed by the UCGIS?' In this paper, we begin to address this question by semantically analyzing 770 published GIS articles in the years following the first publication of UCGIS research priorities.

Table 1. UCGIS Research Priorities. The first year of publication is in parenthesis.

Spatial data acquisition and integration. (1996)
Distributed computing. (1996)
Extensions to geographic presentations. (1996)
Cognition of geographic information. (1996)
Interoperability of geographic information. (1996)
Scale. (1996)
Spatial analysis in a GIS environment. (1996)
The future of spatial information infrastructure. (1996)
Uncertainty in geographic information and GIS-based analyses. (1996)
GIS and society. (1996)
Geospatial data mining and knowledge Discovery. (2000)
Ontological foundations for GIS. (2000)
Geographic visualization. (2000)
Remotely acquired data and information in GIScience. (2000)

1.2 Bibliographic Analysis

The history of using publication data to derive information on academic research begins with bibliographic coupling. The concept of bibliographic coupling, where one measures the amount that different papers cite the same sources, originated in 1963 (Kessler, 1963). The initial work drew few conclusions, but did present a methodology that would later be expanded (Small, 1973; Small, 2003). Co-citation analysis is an outgrowth of bibliographic coupling. Bibliographic coupling compares how often two papers are cited in unison, while co-citation analysis is used for multiple papers cited by multiple additional authors (Small, 1973). The first co-citation map appeared in 1981 from White and Griffith (1981).

The emphasis of co-citation analysis is to determine the subject similarity and association of key ideas in a field, also known as the specialty structure of science (Small, 1973.) A further relationship is established in the social structure of science, which can be determined by the co-authorship linkages in a science authorship network. An authorship network is a social network where vertices are representative of authors and linkages are co-authorship status on one or more journal articles.

Co-authorship networks can be interpreted as structural representations of the collaborative nature of scientific research. The network structure of scientific collaborations has become an interest of great study, in part, because the data are easily available and relatively complete. Using standard analysis tools, several measures can be extracted from the social network structure of co-authorship networks (Newman, 2000).

1.3 Knowledge Domains and Latent Semantic Analysis

Research in the field of information visualization has also examined the issues of innovation and knowledge diffusion through citation analysis and semantic analysis. Knowledge domain visualization is the study of “the dynamic, self-organized, and emergent complex intellectual system that underlies a topical theme, a field of study, a discipline, or an emergent science” (Chen, 2006). There are three key procedures in producing a visualization: extracting the salient structures from a set of data, detecting abrupt changes and emerging trends, and creating a visualization to coherently represent a set of complex information.

Knowledge domain visualization also uses co-citation networks; but unlike social network analysis, the underlying structure of the network and its particular properties are largely ignored. Instead, network analysis tools are replaced with several techniques such as latent semantic analysis and pathfinder network analysis (Chen, 1999).

Latent semantic analysis (LSA, also known as LSI, latent semantic indexing) computes the singular value decomposition matrices of a matrix showing the occurrences of keywords or phrases as columns by the sources (books, journal articles) as rows.. A binary function (0 or 1) can be substituted for the number of occurrences if only the appearance of a work or phrase is to be considered. Singular value decomposition (or SVD) is a technique in matrix computations that divides a matrix into three orthogonal sub-matrices, one of which will be a truncated singular value matrix containing the factors that explain the variance across the rows.

The advantage of using LSA is that related pairings can be determined even when exact words or phrases do not match (Deerwester et al., 1990). This is done by representing the SVD results geographically in an n-dimensional space. The dot-product of two vectors represents their similarity. A network can then be built based on the measurement of similarities, which can then be used to visually represent the connections among journal articles.

Chen (1999; 2006) has provided multiple examples of how co-citation analysis and document similarity can be used to create detail-rich, graphical models of knowledge domains. His work examines the Invisible Colleges, or scientists in a specialty group that may collaborate outside of normal geographic boundaries. In one work, he digested the SCI (Scientific Citation Index) to find journal articles related to string theory in particle physics. Among 150,000 articles, he selected only those that had been cited 35 or more times. The resulting documents were then weighted based on time, since more recent documents may not had as long as a time to become cited as older documents. The result created a graphic that visualized, in a simple but profound way, the knowledge relationships of research in the field of particle physics (Chen, 1999).

2 Methodology

2.1 Data Collection, Processing, and Verification

The principle sets of data for this study include the complete text of the GIS articles from the journals *Transactions in GIS*, *International Journal of GIS*, *GeoInformatica*, and related papers

that appear in the *Annals of the Association of American Geographers* and the *Professional Geographer*. In total, there are 770 articles from the year 1997 to 2007. The data collected includes the date of publication, the names of the authors and their affiliations at the time of the writing. Editorials, book reviews, and other non-research articles were not included in this study. In some cases, articles were excluded due to no source text available. Tables 2, 3, and 4 provide a summary of the distribution of sources in this study.

Table 2. Article frequency by year.

1997	37	1998	38	1999	77	2000	83
2001	71	2002	73	2003	72	2004	74
2005	97	2006	91	2007	57		

Table 3. Article Summary.

Total articles	770
Total distinct first authors	647
Total locations	339
Total journals	5

Table 4. Article count by journal.

International Journal of Geographical Information Science	325
Transactions in GIS	233
GeoInformatica	137
Annals of the Association of American Geographers	40
Professional Geographer	35

To extract author names and affiliations, a small perl program was written for each journal. Accuracy can be checked by alphabetizing the list of author names and affiliations. Common discrepancies are capitalization choices in names and small differences in the naming of universities or colleges. These have been found and corrected to match as necessary. Authors with similar names and affiliation have been assumed to be the same author. Authors with similar names but different affiliations have been considered as separate researchers. The primary author's affiliation location (ie, university) was geocoded through Google Maps.

2.2 Latent Semantic Analysis

Latent semantic analysis begins by creating a textmatrix (Deerwester, et al., 1990). The textmatrix has a list of journal articles in the row-space and the frequency of word occurrences (word counts) within the articles in the column-space. The LSA module for the R programming

language can automatically extract every word combination from the journal articles.

A total of 602 common words, such as articles, pronouns, common verbs, and prepositions, are ignored, as well as most proper names. Words that connote the same meaning but are spelled differently, such as “color” and “colour”, or pluralizations and their singular counterparts, “color” and “colors”, were included in the same column as essentially the same word. Words are broken down into word-stems, or roots of the word. This removes pluralization and adjusts for different verb forms and tenses.

After creating a matrix of journal articles to words, a singular value decomposition (SVD) is performed to create three matrices: $X = T_0 \cdot S_0 \cdot D_0'$, where T_0 and D_0 have orthogonal columns, and S_0 is a diagonal matrix of $r \times r$, where r is the rank of X . The next step is to compute an approximate matrix χ that is generated from the largest k values of S_0 , T_0 , and D_0' into T , S , and D' . This matrix χ contains the independent associational structures in the matrix with the noise removed. The SVD can be interpreted geometrically. The result of the SVD is a k -dimensional vector representing the location of each keyword and journal article.

2.3 Correlation

After generating a latent semantic index, we can run Pearson product-moment correlations on each column-by-column or row-by-row. These provide a similarity index between each object pair. For each UCGIS keyword, we ran correlations with every other term to find words that are highly correlated to the original keyword. In Table 5, we have listed the word-stems that are most highly correlated to the word-stem "mobil," used here to connote the word "mobile."

Table 5. Word-stems derived with a high (> .5) correlation to "mobil" (Mobile computing).

Word-stem	Correlation to "mobil"
mobil	1
phone	0.95
wireless	0.84
hyperlink	0.81
devic	0.72
redirect	0.71
alert	0.68
hypertext	0.64
widespread	0.62
journey	0.56
schedul	0.54
wayfind	0.54
envelop	0.53

3 Research Trends in Geographic Information Science

3.1 Latent Semantic Analysis: Correlating Similarity to Distance

The correlation of the latent semantic analysis of the article textmatrix yields a similarity index between each article. Below is a graph showing the similarity index between articles versus the distance between the authors' locations.

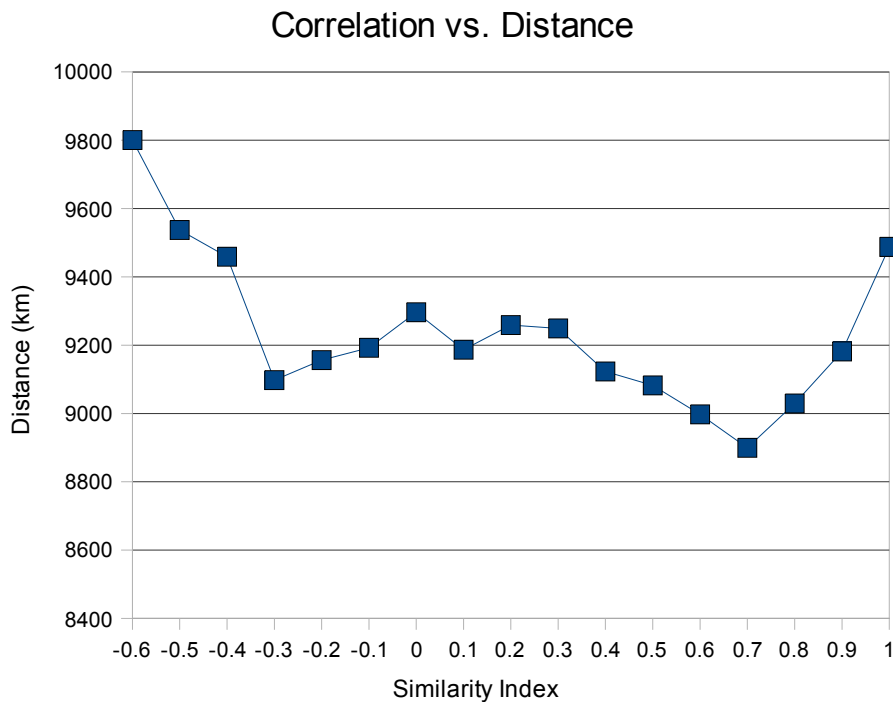


Illustration 1. Similarity Index versus Distance (km)

Geographic distance is not linearly related to an article's similarity. Fundamental research in GIS is geographically diverse, with primary research located in North America, Europe, and Australia.

3.2 The Research Priority Network.

In Section 2.3, we found the terms that were highly correlated with the UCGIS subject keywords. To show how the subject keywords are related, we have correlated each keyword with each other keyword. In Illustration 2, we link subjects that have a .5 Pearson correlation or higher. The network is a conceptual framework to demonstrate how the research priorities are thus related in the GIS knowledge domain.

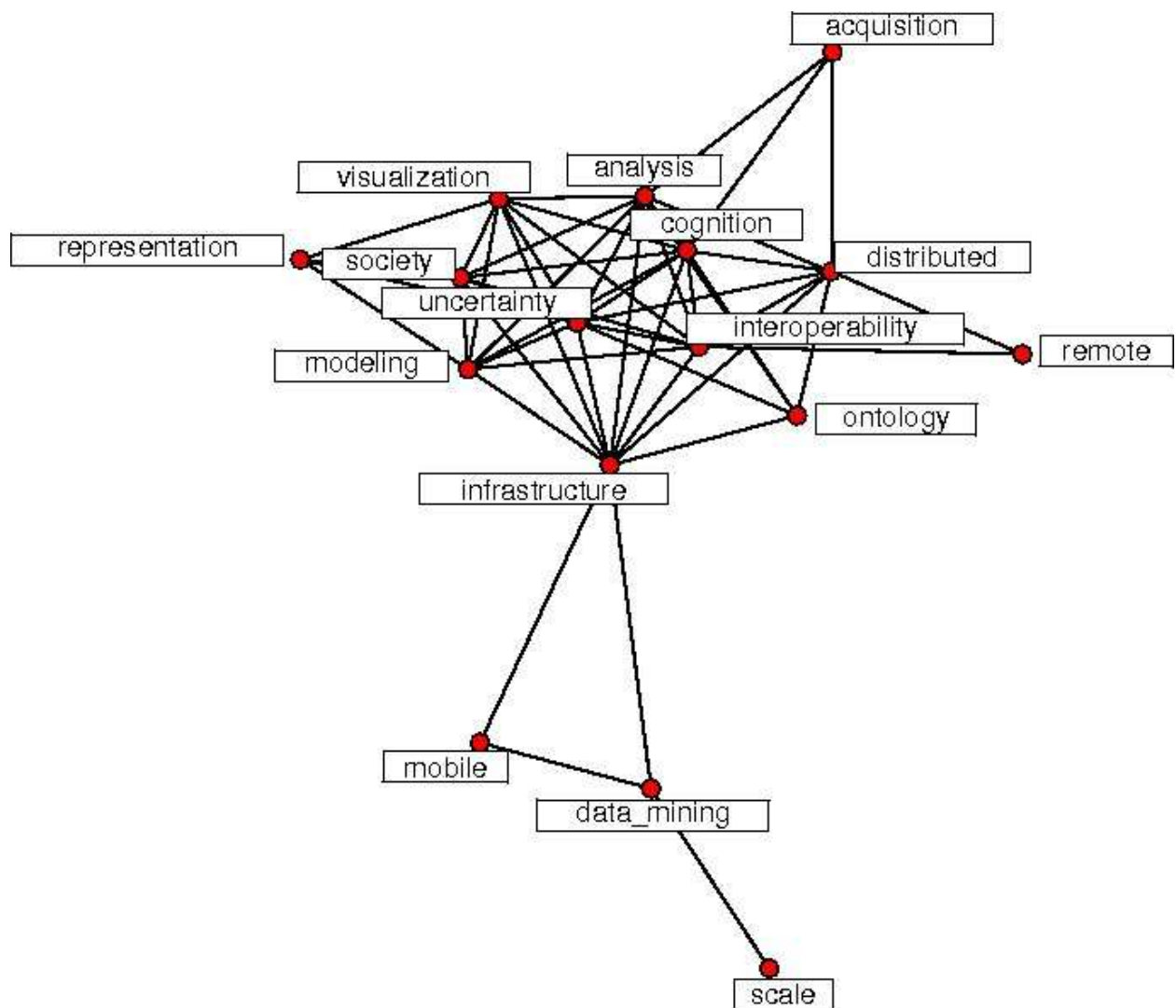


Illustration 2. The knowledge domain network of UCGIS research priorities.

3.3 Priorities Over Time

With the original textmatrix substituted for a year-by-keyword matrix, we can quantify the year-by-year research results against the UCGIS research priorities after running latent semantic analysis. (See Table 6). From this, we can see several trends in GIS research. Some areas, such as modeling, representation, and acquisition, remain highly active from year-to-year. Research in infrastructure has become less active over time. Mobile computing research trends upward in the years 2006 and 2007.

Table 6. Research quantified: results from LSA of the year-by-year research priorities.

	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
modeling	16.9	17.9	17.6	19.2	17.1	18.2	16.7	17.8	17.7	15.9	15.4
data mining	9	7.9	8.2	6.5	8.8	7.6	9.2	8	8.1	10.1	10.6
ontology	7.8	6.1	6.6	3.9	7.6	5.6	8.2	6.3	6.5	9.6	10.4
acquisition	9.1	8.6	8.7	8	9	8.5	9.1	8.7	8.7	9.5	9.7
visualization	7.3	6.4	6.7	5.2	7.2	6.2	7.5	6.5	6.6	8.3	8.7
representation	8.6	8.9	8.8	9.3	8.6	9	8.5	8.9	8.8	8.2	8.1
society	7.7	7.5	7.6	7.3	7.7	7.5	7.7	7.5	7.6	7.9	8
analysis	5.8	6.3	6.2	6.9	5.9	6.4	5.7	6.2	6.2	5.3	5.1
infrastructure	7.8	9.7	9.2	12.2	8	10.3	7.4	9.5	9.3	5.7	4.8
interoperability	3.3	2.7	2.9	1.8	3.3	2.5	3.6	2.8	2.8	4	4.4
cognition	2.6	2.1	2.2	1.6	2.5	2	2.7	2.2	2.2	3.1	3.3
mobile	1.7	0.8	1	0	1.5	0.5	1.8	0.9	1	2.6	3
remote	3.6	4.2	4	4.9	3.7	4.3	3.5	4.1	4	2.9	2.7
distributed	3.2	3.6	3.5	4.1	3.2	3.7	3.1	3.6	3.5	2.8	2.6
scale	3.5	4.4	4.1	5.5	3.6	4.6	3.3	4.3	4.2	2.6	2.2
uncertainty	2.2	2.9	2.7	3.8	2.3	3.1	2.1	2.8	2.7	1.5	1.1

3.4 Locational Analysis

By modifying the original article-word textmatrix, we can create a location-word textmatrix and perform the latent semantic analysis as well. Doing so, we can find which locations have a high correlation with a particular research path. It is possible that locations with a high correlation to a subject area would be key innovative sites for that field. Illustrations 3,4 and 5 show primary research locations for some of the subject keywords.

With this, this paper answers the question of where the key centers of GIS research are. Further research may consider the ties between locations or the geographic network of each subfield. Knowing the relationship of research locations in each subfield could assist in identifying sources of new research priorities.



Illustration 3. Primary research subject locations in Europe.

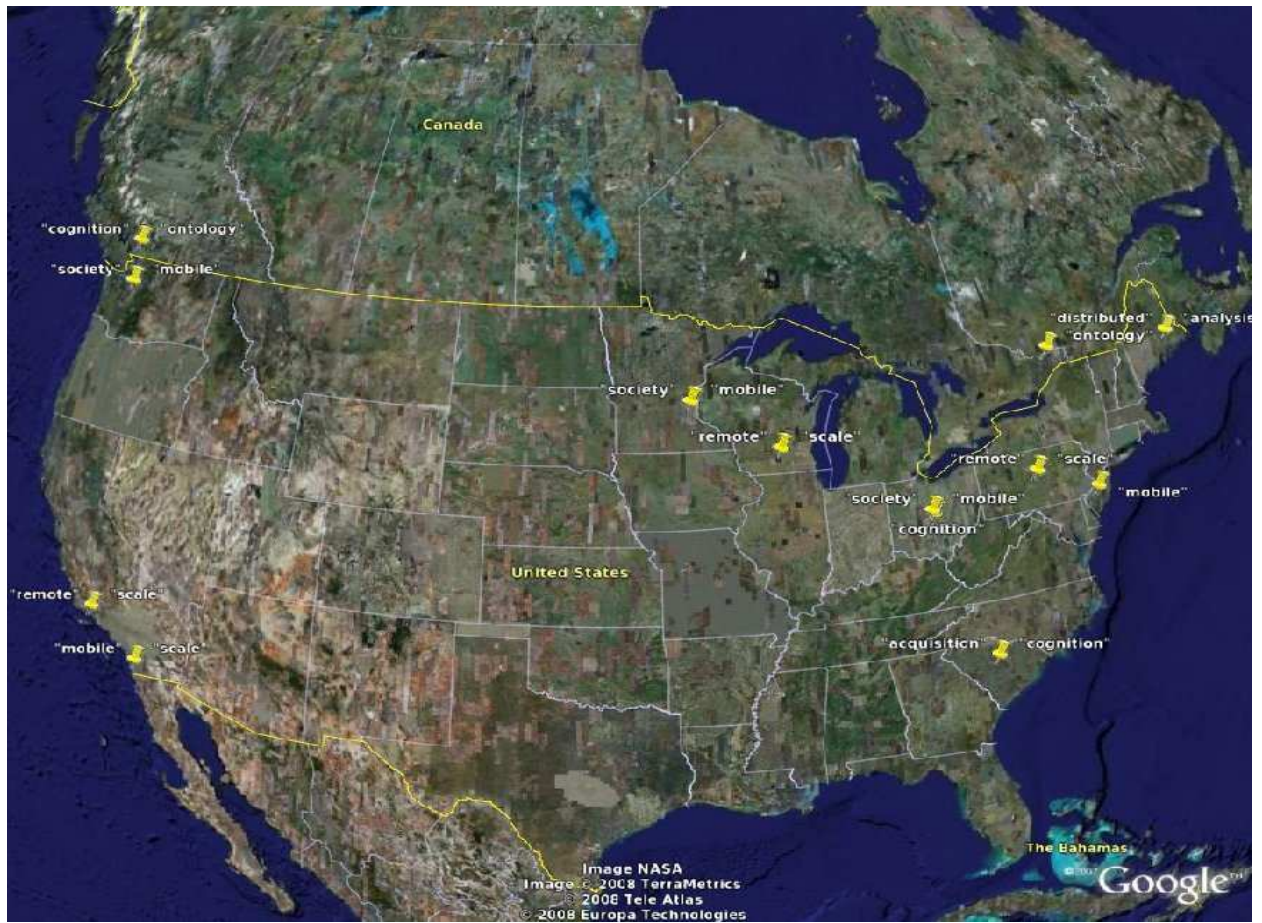


Illustration 4. Primary research subject locations in North America.

Cognition	Data Mining	Distributed Computing	Spatial Data Infrastructure
Burnaby, BC	Beijing, China	Canberra, Australia	Columbus, OH
Columbia, SC	Burnaby, BC	Edinburgh, UK	Enschede, Netherlands
Columbus, OH	Hong Kong, China	Hong Kong, China	Hong Kong, China
Enschede, Netherlands	Ispra, Italy	Leuven, Belgium	Leeds, UK
Hong Kong, China	London, UK	London, UK	London, UK
Leeds, UK	Madison, WI	Madison, WI	Melbourne, Australia
Madison, WI	Orono, ME	Melbourne, Australia	San Diego, CA
Melbourne, Australia	Santa Barbara, CA	Orono, ME	Santa Barbara, CA
Munster, Germany	State College, PA	San Diego, CA	Seattle, WA
State College, PA	Vienna, Austria	Santa Barbara, CA	State College, PA
Remote Computing	Representation	Scale	GIS and Society
Canberra, Australia	Hong Kong, China	Canberra, Australia	Camden, NJ
Edinburgh, UK	Ispra, Italy	Edinburgh, UK	Columbus, OH
Hong Kong, China	Leeds, UK	Enschede, Netherlands	Enschede, Netherlands
London, UK	London, UK	Hong Kong, China	London, UK
Madison, WI	Madison, WI	London, UK	Melbourne, Australia
Melbourne, Australia	Orono, ME	Madison, WI	Minneapolis, MN
Orono, ME	Richmond, BC	Melbourne, Australia	San Diego, CA
Santa Barbara, CA	Santa Barbara, CA	San Diego, CA	Santa Barbara, CA
State College, PA	State College, PA	Santa Barbara, CA	Seattle, WA
Vienna, Austria	Vienna, Austria	State College, PA	State College, PA
Interoperability	Mobile Computing	Uncertainty	Visualization
Burnaby, BC	Camden, NJ	Canberra, Australia	Columbia, SC
Enschede, Netherlands	Columbus, OH	Edinburgh, UK	Columbus, OH
Hong Kong, China	Enschede, Netherlands	Hong Kong, China	Enschede, Netherlands
Leeds, UK	London, UK	Leuven, Belgium	Hong Kong, China
London, UK	Melbourne, Australia	London, UK	London, UK
Melbourne, Australia	Minneapolis, MN	Orono, ME	Madison, WI
Munster, Germany	San Diego, CA	Richmond, BC	Melbourne, Australia
Orono, ME	Santa Barbara, CA	Santa Barbara, CA	Santa Barbara, CA
Santa Barbara, CA	Seattle, WA	Vienna, Austria	State College, PA
State College, PA	Southampton, UK	Zurich, Switzerland	Stuttgart, Germany

Illustration 5. The Ten Highest Correlated Locations Per Research Area.

4 Conclusions

4.1 The Impact of the UCGIS Research Priorities

In measuring the research related to the UCGIS research priorities, we can determine the baseline of research in the geographic information science field. As the priorities have changed over the eleven year span of 1997-2007, so has the research changed as well. Trends in the research can be enumerated, providing guidance on current research priorities in the field. Given these results, the UCGIS can reconsider which research priorities need to be adjusted, encouraged, or removed.

It should be noted that the results do not speak to the quality of the research pursued. They do,

however, provide the UCGIS with information on how research has changed, where different types of research are being performed, and how the knowledge domain is structured internally.

4.2 Future Research

Semantic analysis provides a method to identify and quantify the relationship among published research. While it may not speak to the quality of the research, it can establish linkages between research, keywords, time, and space. This article has demonstrated several methods of analyzing the research changes over time (from 1997 to 2007) and over space. The analysis has identified words closely related to the subject keywords. Future research may explore how semantic alternatives are used in different locations.

Previous academic research has focused on the citation network as the structure of scientific connectivity (Small, 1973; White and Griffith, 1981; Newman, 2000; Newman, 2001). Citation networks are limited in that all citation values are binary: 1 for a work that is cited, 0 for a work that is not cited. Not all citations can be assumed to be of equal value, however. Latent semantic analysis could be used to find the similarity indices between journal articles and other published scholarly research. A similarity index may be a suitable replacement for the binary citation values.

References

- Barabasi, A.L., H. Jeong, Z. Neda, E. Ravasz, A. Schubert, T. Vicsek. 2002. Evolution of the social network of scientific collaborations. *Physica A*: 590-614.
- Borner, Katy. Chaomei Chen, Kevin W. Boyack. 2003. Visualizing knowledge domains. *Annual Review of Information Science and Technology* 37:
- Chen, Chaomei. 1998. Generalised similarity analysis and pathfinder network scaling. *Interacting with Computers* 10: 107-128.
- Chen, Chaomei. 1999. Visualizing semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management* 35: 401-420.
- Chen, Chaomei. Jansa Kuljis, Ray J. Paul. 2001. Visualizing latent domain knowledge. *IEEE Transactions on Systems, Man, and Cybernetics: Part C: Applications and Reviews* 31(4): 518-529.
- Chen, Chaomei. Jian Zhang, Weizhong Zhu, Michael Vogeley. 2007. Delineating the citation impact of scientific discoveries. *JCDL 2007, Vancouver, British Columbia, Canada*.
- Chen, Chaomei. Weizhong Zhu, Brian Tomaszewski, Alan MacEachren. 2007. Tracing conceptual and geospatial diffusion of knowledge. *Online Communities and Social Computing* 265-274.
- Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science* 41 (6): 391-407.
- Elmes, Gregory. 2005. Guest Editorial: The University Consortium for Geographic Information Science: Shaping the Future at Ten Years. *Transactions in GIS* 9(3): 273-276.
- Kessler, M. M. 1963. Bibliographic coupling between scientific papers. *American*

- Documentation* 14(1): 10-25.
- Lewison, Grant. Isla Rippon, Steven Wooding. 2005. Downstream influence: tracking knowledge diffusion through citations. *Research Evaluation* 1(1): 5-14.
- Newman, M. E. J. 2000. Who is the best connected scientist? A study of scientific coauthorship networks. *Physics Review E* 64:
- Newman, M. E. J. 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Science* 98: 404-424.
- Small, Henry. 1973. Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science* 24(4): 265-269.
- Small, Henry. 2003. Paradigms, citations, and maps of science: a personal history. *Journal of the American Society for Information Science and Technology* 54(5): 394-399.
- University Consortium of Geographic Information Science. 1996. Research priorities for geographic information science. *Cartography and Geographic Information Systems* 23(3): 115-127.
- White, Howard D. and Griffith, B. C. 1981. A cocitation map of authors in judgment and decision research. *Journal for the American Society of Information Science* 32: 163-172.
- White, Howard D. 2003. Pathfinder networks and author cocitation analysis: a remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology* 54(5): 423-434.