SEARCHING FOR PATTERNS IN UNSTRUCTURED TEXT

David A. Thomas

University of Pretoria

Worldwide, scholars rely on information technologies to search for and retrieve publications and data pertinent to their research interests. But what if a search identifies hundreds of documents, any of which might be vital or irrelevant? Automated text analysis technologies like Leximancer may be used to characterize large sets of unstructured data. This paper applies this approach to 138 scholarly papers dealing with teaching and learning geometry.

INTRODUCTION

According to Smith and Humphreys (2006), there are several reasons why one would want an automated system for content analysis of documents. Researchers are subject to influences that they are unable to report (Nisbett & Wilson, 1977) which may lead to subjectivity in data analysis and the interpretation of findings. Limiting researcher subjectivity often involves extensive investments of time and money to address inter-rater reliability and other sources of bias. One goal of automated text analysis is to reduce this cost and to allow more rapid and frequent analysis and reanalysis of text. A related goal is to facilitate the analysis of massive document sets and to do so unfettered by *a priori* assumptions or theoretical frameworks used by the researcher, consciously or unconsciously, as a scaffold for the identification of concepts and themes in the data (Zimitat, 2006).

Since textual analysis technologies operate directly on words (as well as other symbols), a rationale for inducing relationships between words is needed. Beeferman observed that words tend to correlate with other words over a certain range within the text stream (Beeferman, Berger, & Lafferty, 1997). Indeed, a word may be defined by its context in usage (Smith and Humphreys, 2006, p. 262; Courtial, 1989; Leydesdorff and Hellsten, 2006; Lee and Jeong 2008).

According to the *Leximancer* (2010) manual, concepts are "collections of words that typically travel together throughout the text". For example, in a document about climate change, the *Leximancer* concept *carbon* might include the keywords *dioxide, carbonate, footprint*, and *sequester*. *Leximancer* weights these terms according to how frequently they occur in sentences containing the concept, compared to how frequently they occur elsewhere. A sentence (or group of sentences) is only tagged as containing a concept if the accumulated evidence (the sum of the weights of the keywords found) is above a set threshold. *Leximancer* induces the definition of each concept (i.e. the set of weighted terms) by noting the co-occurrence of words within a "sliding window" as it scans blocks of text a few sentences at a time (Liesch, Hakanson, McGaughey, Middleton, and Cretchley 2011 and Grimbeek, Bryer**,** Davies, & Bartlett 2005). These data are used to make two determinations: (i) the most frequently used concepts within a body of text; and more importantly, (ii) the relationships between these concepts (e.g., the co-occurrence between concepts). During the learning process, words highly relevant to the seed are continuously updated, and eventually form a thesaurus of terms for each concept.

This approach to textual analysis is well founded. The repeated co-occurrence of words within a document or a set of documents is a widely accepted principle (Clandinin and Connelly 2000; Sowa, 2000). Concepts consisting of co-occurring words do reflect categories that carry meaning (Osgood et al., 1957; Leydesdorff and Hellsten, 2006). For

example, a textual analysis of this paper would discover that the words “text” and “analysis” co-occur frequently.

INVESTIGATING PROFESSIONAL KNOWLEDGE BASES

Leximancer has been used in a number of studies to analyze the content and evolution of professional knowledge bases embodied in journal publications and conference proceedings. Indulska and Recker (2008) studied the progress of design science research in Information Systems (IS) as seen in papers published at the main five AIS-sponsored IS conferences, namely ACIS, AMCIS, ECIS, ICIS and PACIS in 2005-2007. Their analyses found that design science is a growing stream of research in Information Systems and that design science research is widespread in the research domains of process, knowledge and information management.

Using a similar approach, Liesch, Hakanson, McGaughey, Middleton, and Cretchley (2011) examined articles published in the Journal of International Business Studies (JIBS) from 1970 until 2008. Their findings suggest that the IB literature has evolved from a formative stage in which macro-environmental contextual factors were of prime consideration to its current position in which the firm, its strategy and its performance, have become the dominant interests.

Sadiq, Yeganeh, and Indulska (2011) used *Leximancer* to investigate the current landscape of data quality research, to create better awareness of synergies between various research communities, and, subsequently, to direct attention towards holistic solutions. Their findings include a taxonomy of data quality problems, identification of the top themes, outlets and main trends in data quality research, as well as a detailed thematic analysis that outlines the overlaps and distinctions between the focus of Information Systems and Computer Science publications.

In this study, *Leximancer* was used to characterize the content of 138 scholarly publications focused on teaching and learning geometry. These papers were identified and downloaded using the *Google Academic* and *Springer* online databases. The following research questions were asked:

1. Which concepts occur with the greatest frequency?
2. Which concepts are closely associated with one another?
3. What are the emergent themes?
4. To what extent do different papers and/or authors emphasize similar concepts?

INTRODUCTION TO LEXIMANCER

According to Smith and Humphreys (2006), *Leximancer* discovers and extracts thesaurus-based concepts directly from the text data. These concepts are then coded into the text (i.e., tags are inserted) using the thesaurus as a classifier. This process employs two stages of co-occurrence information extraction—*semantic* and *relational*—using a different algorithm for each stage. Clusters of co-occurring concepts are then aggregated into themes. The algorithms used are statistical, but they employ nonlinear dynamics and machine learning. The resulting asymmetric concept co-occurrence information is then used to generate a concept map

Leximancer processes text in a series of eight stages (See Figure 1): Load Data, Pre-Process, Concept Seeds Identification, Edit Emergent Concept Seeds, Develop Concept Thesaurus, Create Compound Concepts, Code Concepts into Text, and Generate Outputs. At each stage, the researcher is prompted to edit the manner in which *Leximancer* processes the data provided. The editing options provided at each of these stages enable the researcher to, in effect, change the size of the sliding window, add or delete documents, merge similar concepts (e.g., student, students, pupils), delete irrelevant or distracting

concepts (e.g., recurring formatting terms like *Figure* and *Table*), propose additional concepts, and so on. All such choices are reversible, so experimentation is both possible and desirable in refining the analysis and presentation graphics. In practice, there is no limit to the number of documents that may be analyzed using *Leximancer*, provided that the computer on which the software is running is adequate to the task.

Three additional options are available at the bottom of the control panel: Concept Map, Insight Dashboard, and Data Exports. An analysis of the data reported in the Insight Dashboard identified 38,338 context blocks and 70 concepts.

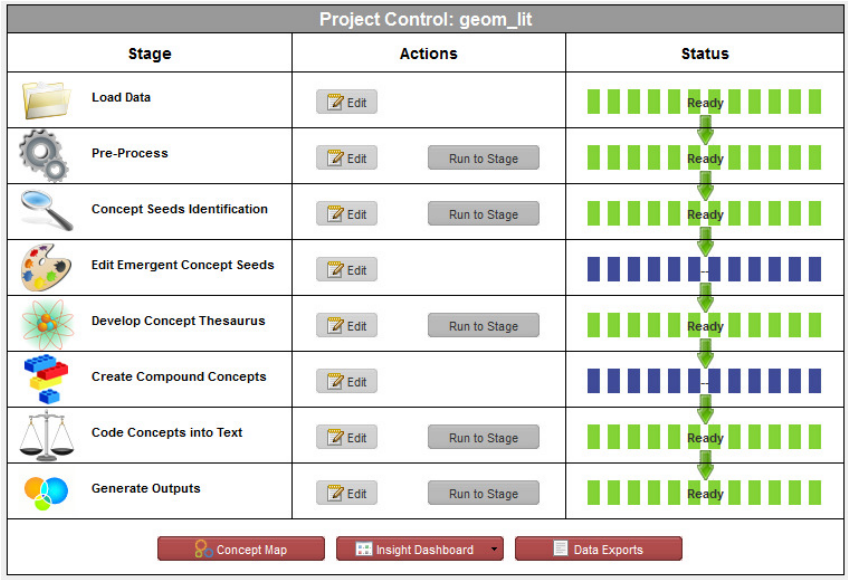

Figure 1 *Leximancer* Control Panel

Figure 2 shows a network of nodes and connecting line segments. In this figure, the 70 nodes near the center of the figure correspond to the 70 concepts discovered. Further from the center are 138 paper nodes identified by first author.

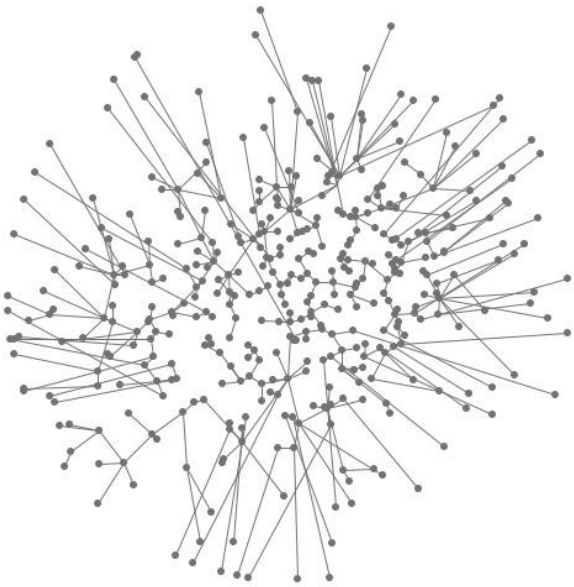

Figure 2 Network of Concepts and Papers

In Figure 3, the concepts discovered by *Leximancer* are displayed adjacent to their respective nodes. Note that the concepts *mathematical, student, teacher, technology, use,* and *learning* appear as large, red dots near the center of the map. By contrast, the concepts *angle* and *triangle* appear as smaller blue dots on the periphery. In *general,*

- Green concept labels represent proper names (such as people or locations). Black concept labels refer to other objects, locations, actions and so on.
- Concept nodes are heat-mapped, in that hot colours (red, orange) denote the most relevant concepts, and cool colours (blue, green), denote the least relevant.
- The brightness of a concept's label reflects its frequency in the text. The brighter a concept label the more often the concept is coded in the text.
- Concepts are contextually clustered on the map. That is, concepts that frequently appear together in the text settle close to one another on the map.

An alternative look at the discovered concepts is seen in Figures 4 and 5. In Figure 4, the concepts are listed in ranked order based on the number of times they occur (e.g., Count) in the set of papers. This statistic is then converted into a Relevance score, computed as the percentage frequency of text segments which are coded with that concept relative to the frequency of the most frequent concept in the list. In that list, we see that the concepts *technology* and *learning* occur, respectively, 2893 and 3449 times. When a concept in the ranked list is clicked, such as *technology*, another screen is displayed (see Figure 5) showing the co-occurrence of the concept *technology* with other discovered concepts, such as *learning*. In that line of the table, we see that *technology* co-occurred with *learning* 580 times. The Likelihood score of 17% may be interpreted as saying that the concept *technology* co-occurs with the concept *learning* 17% of the time. Clicking on the magnifying glass icon to the left of the *learning* concept opens a window in which each of those 580 co-occurrences is displayed as seen in the original paper.

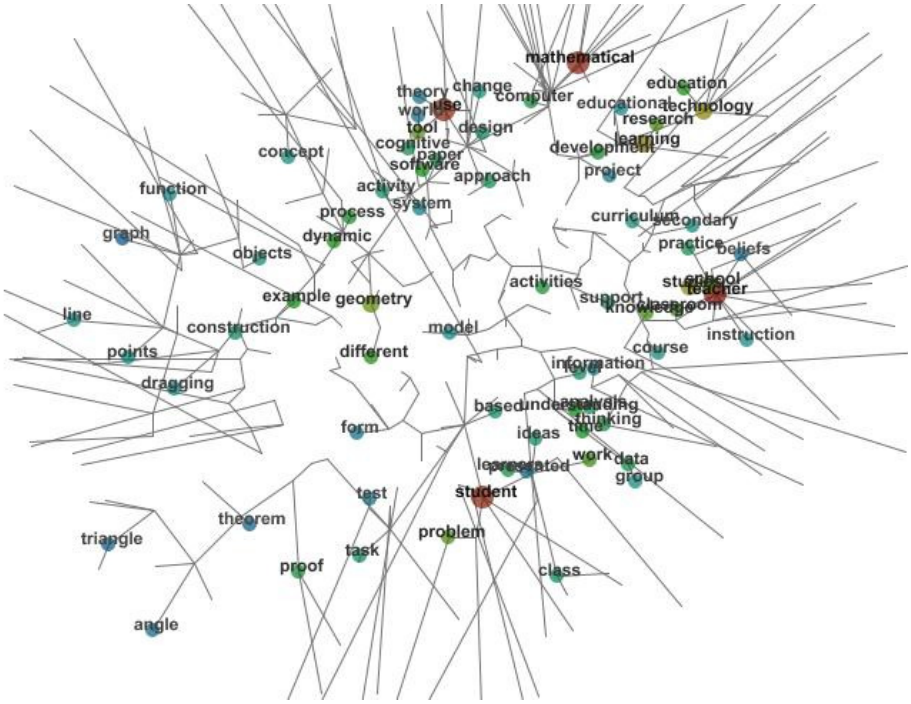

Figure 3 Concept Map

Themes | **Concepts** | Thesaurus | Pathway | Query | Summary

Ranked Concepts Export

Word-Like	Count	Relevance
teacher	9414	45%
mathematical	9229	44%
student	9220	44%
use	7449	36%
learning	3449	17%
technology	2893	14%
studies	2797	13%
problem	2467	12%
geometry	2374	11%
school	2335	11%
tool	2151	10%
research	2119	10%
work	1956	09%
knowledge	1952	09%
different	1837	09%
classroom	1755	08%
education	1638	08%
example	1554	07%
understanding	1455	07%
process	1440	07%
proof	1361	07%
time	1360	07%
dynamic	1339	06%
software	1275	06%
development	1224	06%
data	1166	06%
class	1163	06%
thinking	1143	05%
task	1133	05%
learners	1113	05%
computer	1099	05%
activities	1095	05%
paper	1064	05%
analysis	1016	05%

Figure 4 Ranked Concepts

Themes | **Concepts** | Thesaurus | Pathway | Query | Summary

Concept: **technology** Ranked View Export

Related Name-Like	Count	Likelihood
Related Word-Like	**Count**	**Likelihood**
educational	138	24%
education	367	22%
beliefs	76	20%
classroom	314	18%
support	141	18%
use	1287	17%
learning	580	17%
practice	159	17%
curriculum	122	17%
instruction	113	16%
information	99	16%
teacher	1465	16%
course	109	15%
computer	159	14%
change	118	14%
design	112	14%
development	170	14%
research	293	14%
mathematical	1225	13%
tool	284	13%
studies	359	13%
school	291	12%
software	157	12%

Figure 5 Co-occurring Concepts

Themes are clusters of concepts. Like concepts, themes are heat-mapped, meaning hot colors (red, orange) denote the most relevant themes, and cool colors (blue, green), denote the least relevant. In this study, *Leximancer* discovered 70 concepts. While each of these may be treated as a theme, most researchers prefer to aggregate or nest these concepts into a small number of overarching themes. Figure 6 lists six such themes. Theme connectivity is the summed co-occurrence counts of each concept within the theme with all available concepts.

Theme	Connectivity	Relevance
mathematical	100%	
student	93%	
teacher	72%	
dynamic	37%	
line	09%	
angle	04%	

Figure 6 Overarching Themes List

In Figure 7, the same themes are superimposed on the network representing the discovered concepts and papers/authors. Figure 8 overlays concepts with themes. In so doing, it juxtaposes small scale (i.e., concepts) and large scale (i.e., themes) analytic elements. Metaphorically, it is at this point in the analysis that we begin to get a glimpse of the *forest* implicit in the corpus of work as well as the *trees* explicit in the individual papers.

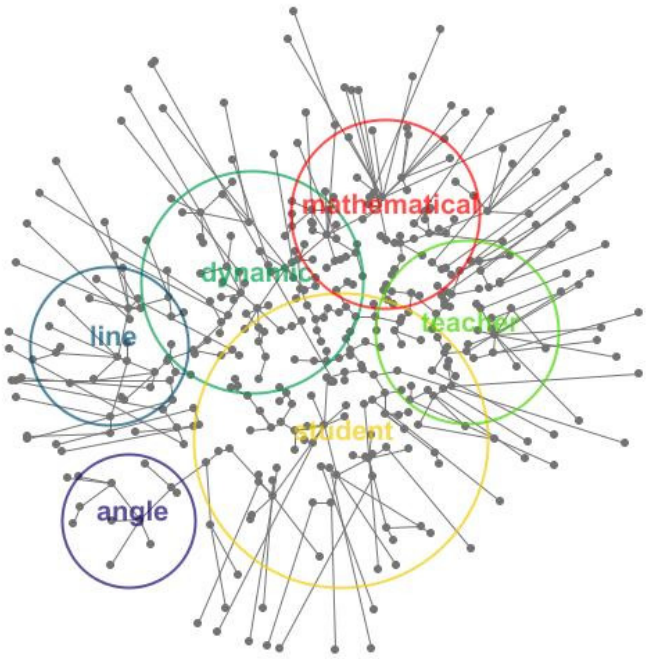

Figure 7 Overarching Themes Superimposed on Node Network

The Thesaurus tab shows which words are associated with any given concept. For instance, in Figure 9, the terms associated with the concept *technology* are listed in the Word column with their respective weights. Clicking on the … icon beside each Word Generates the text passages from which the Words were drawn (See Figure 10). The “Iterations: 7” notation indicates that 7 iterations of learning (i.e., reading and analyzing the text) were used to generate a stable set of concept definitions.

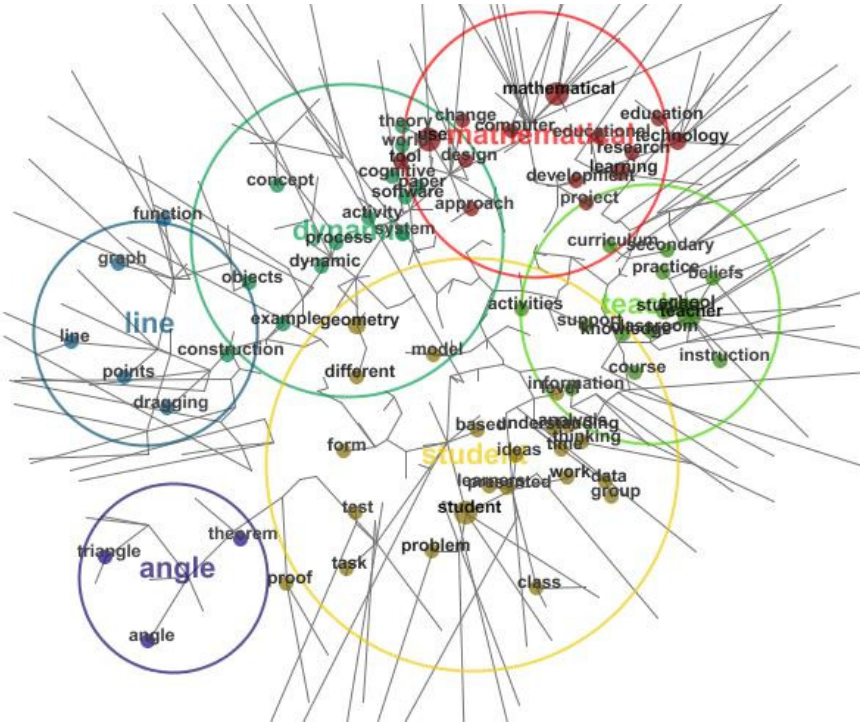

Figure 8 Concepts Superimposed on Major Themes

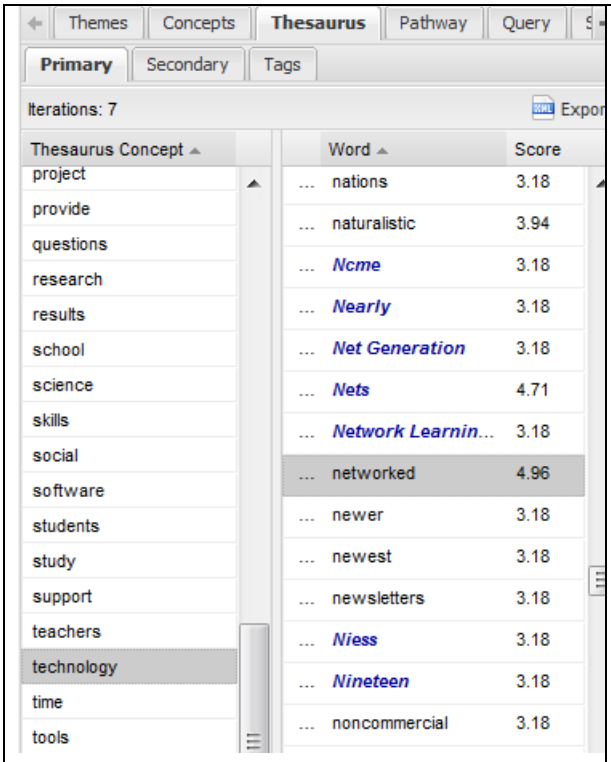

Figure 9 Thesaurus Word Definitions

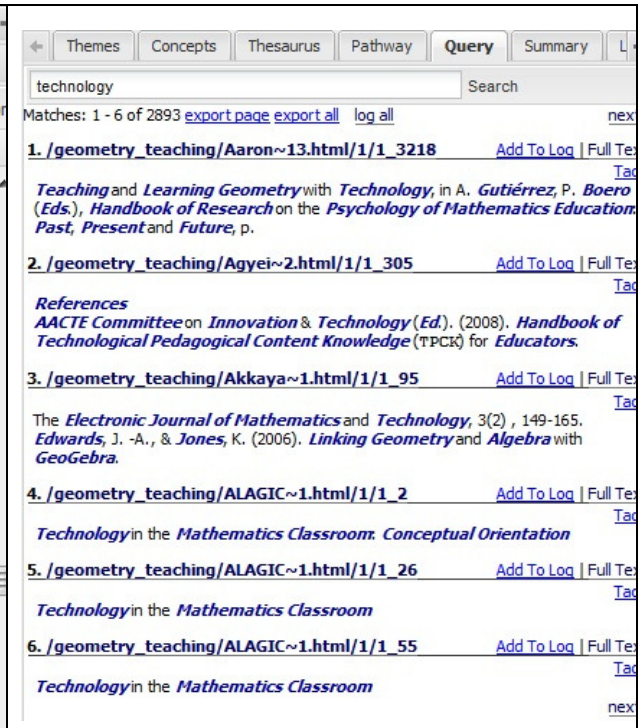

Figure 10 Thesaurus Source Texts

In Figure 11, each paper is positioned relative to the emergent major themes using the names of their first authors. As with concepts, papers emphasizing similar concepts are positioned closer to one another than papers with little in common.

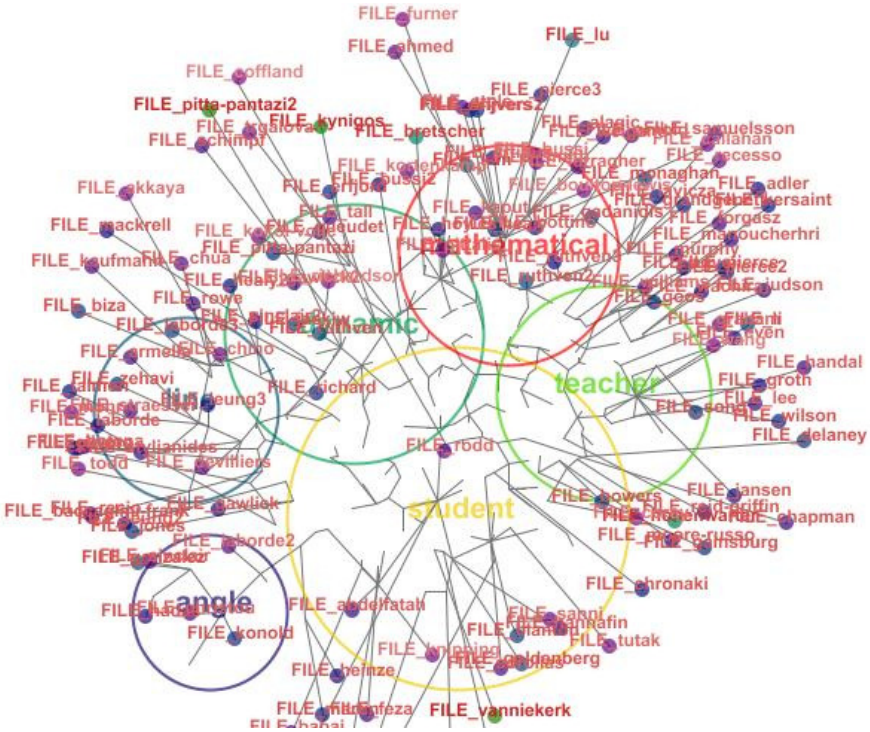

Figure 11 Themes and Papers by First Author

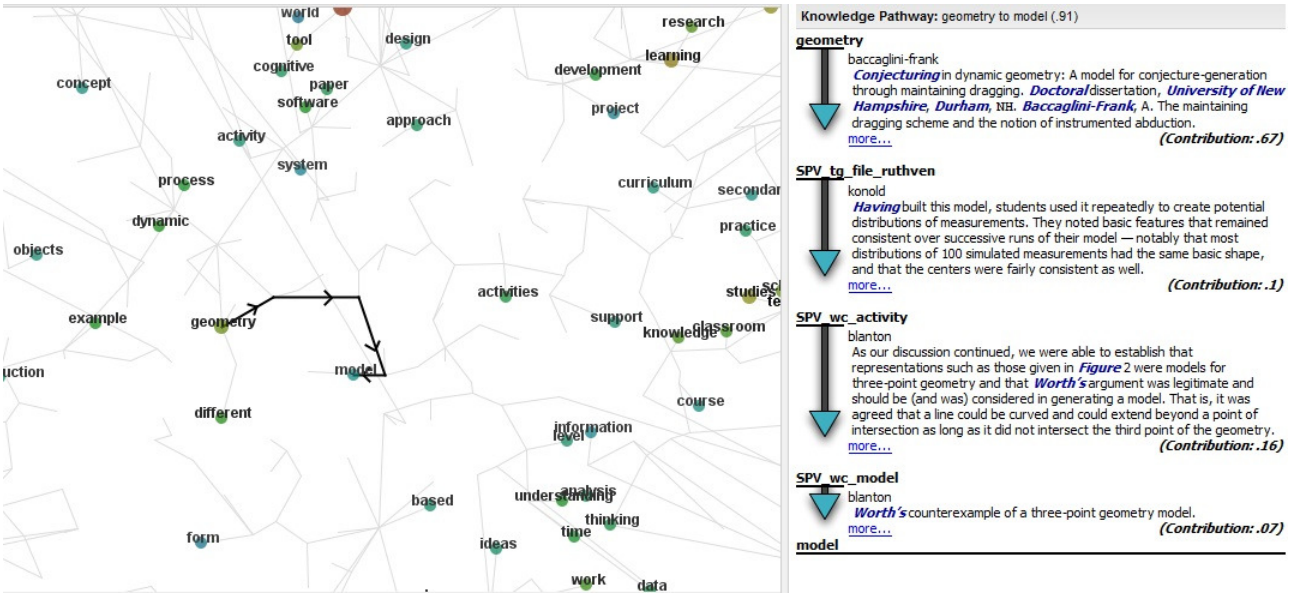

Figure 12 Concept Associations

In Figure 12, the *distance* between the concepts *geometry* and *model* is mapped using a set of segments described in the Knowledge Pathway text box to the right. A measure of their connectedness (0 - 1) is computed, with the indicated value 0.91 representing a strong connection.

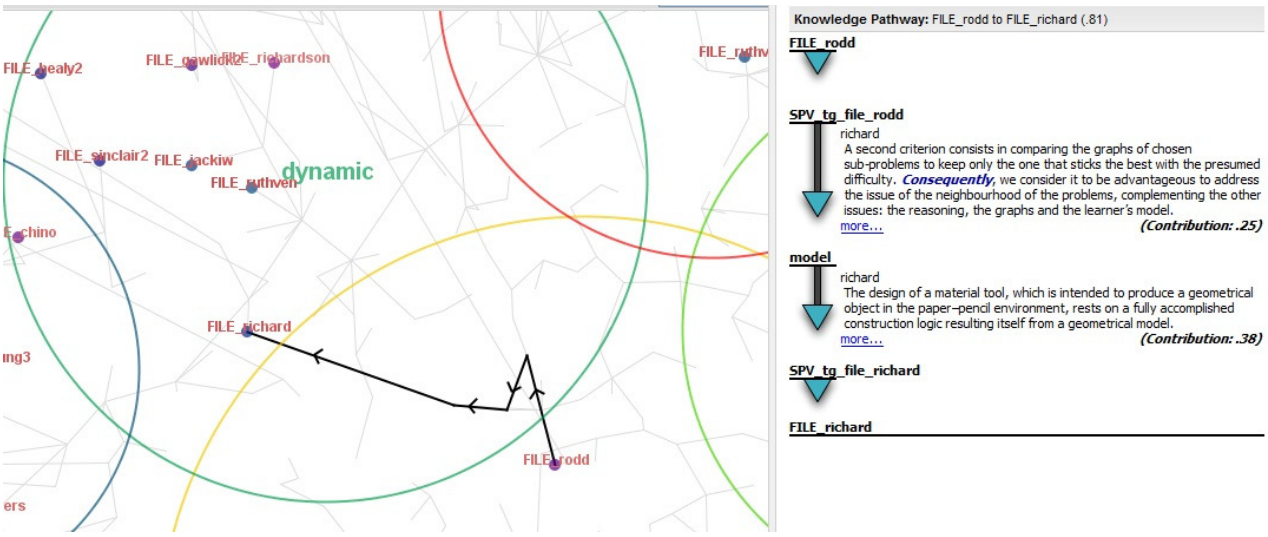

Figure 13 Paper Associations

A similar Knowledge Pathway is presented in Figure 13 representing the connection between two papers. The value 0.81 indicates that these two papers share much in common.

RESEARCH FINDINGS

Analysis of the data provided the following insights:

1. Which concepts occur with the greatest frequency? Teacher, mathematical, student, use, learning technology, studies, problem, geometry, school.

2. Which concepts are closely associated with one another? In Figure 3, related concepts appear near one another in the Concept Map (e.g., education, educational technology, research).
3. What are the emergent themes? The themes mathematical, student, teacher, dynamic, line, and angle provide a useful clustering of concepts (see Figure 8).
4. To what extent do different papers and/or authors emphasize similar concepts? As seen in Figure 11, different papers may be closely or only remotely associated.

These findings are preliminary in nature and incomplete. A detailed summary of the findings implicit in the Leximancer analysis is beyond the scope of this paper. The following discussion offers some speculations on the potential value of automated text analysis as it relates to scholarship and academic advising.

DISCUSSION

We now turn to the potential value of content analysis in a variety of academic contexts, including individual scholarship, collaborative research, and graduate student advising. Admittedly, many of these thoughts are speculative in nature. But the challenges they address are all too real.

Most scholars are participating and contributing members of scholarly communities (e.g., academic departments, government laboratories, museums, archives) and/or associations (e.g., American Education Research Association, American Mathematical Society). One of the functions of professional communities and associations is to sponsor, structure, and moderate an on-going dialogue among members. Meeting and publications provide mechanisms for vetting research findings, albeit without making any guarantees beyond procedural matters. Nevertheless, this approach continues to motivate and facilitate the development and testing of new ideas, so many new ideas that it is all but impossible for modern scholars to “keep up” unaided by sophisticated information technologies.

This paper suggests how *Leximancer* might be used to review, not dozens, but hundreds or even thousands of documents, be they scholarly papers, government reports, interview transcripts, or any other textual material. Since most scholars are, to some degree, specialists, their work is situated within a specialized literature typically published in professional journals, conference proceedings, monographs, books, and online forums. Traditionally, scholars acquired and managed the knowledge base embodied in such hardcopy publications using bookshelves, file cabinets, cross-referencing materials in hardcopy and/or electronic formats, and the most fluid of all media, memory. With automated content analysis such as that provided by *Leximancer*, sophisticated knowledge management is now possible on a scale that has, until recently, been available only to information management specialists. For instance,

- Individual scholars can create digital folders of seminal documents in their research domains (including raw data, reports, journal articles, books, proceedings, dissertations, and works in progress) and analyse the entire corpus of work in terms of concepts, themes, authors, and their connections. In this approach, both the *forest* and its *boundaries* can be contemplated holistically.
- Teams of scholars working within or across disciplines can more readily share their respective interests and experiences, juxtapose their potential contributions, and identify challenges on the boundaries and at the intersections of their respective fields. Using the same approach to investigate funding opportunities, teams of researchers can become better informed and smarter players in grants and contracts competitions. For instance, funding sources typically publish RFPs for recent, current, and anticipated grant competitions on public websites. Many private foundations and governmental agencies also publish abstracts of funded grants and related PI contact information. A content analysis of this information and related research publications can be of significant

strategic value when deciding how and when to invest faculty time in the development of grant proposals.

- When approached by graduate students struggling with their research proposals, thesis and dissertation advisors can say with assurance, “Here is what matters most in this knowledge domain. Start here.” When a student presents his/her research proposal for review, the proposal can be temporarily added to the knowledgebase and analysed relative to its content. That analysis can be used to motivate and facilitate discussions essential to the mentor-student relationship and to address common issues such as, “How are your research questions and theoretical framework grounded in the literature? How does your literature review compare to those of published papers? Are your research design, sampling, and data analysis procedures consistent with accepted practice? What have you included that is not in the knowledge base? How do you justify doing so?” And so on. Using the Bayes Factor (Wikipedia, 2011) statistic, it is even possible to formulate and test hypotheses relative to these and other questions addressing similarities and differences between the student’s proposal and the knowledgebase as a whole. Here is a bridge between qualitative and quantitative methodologies on which their respective proponents might meet.
- University administrators are constantly seeking better information relative to the effectiveness of student recruiting, transition to campus life, retention, achievement, and employment. For the most part, these data are gathered using closed response format items (i.e., true-false, multiple choice, Likert scale) and university student records. To date, no golden needle has been found in that haystack. Using content analysis methods, open response items, interview transcripts, and samples of student work can be “unpacked” to find the real roots of success/failure and satisfaction/disappointment for students, faculty, and staff in their respective roles. The problem has always how to reliably and rapidly analyse tens of thousands of documents. With automated content analysis, the issue is how to motivate, facilitate, and find wisdom in authentic student comments and recommendations.

Learning content analysis methodologies and technologies is neither a cheap nor a simple venture. A desktop copy of *Leximancer* at academic rates is AD$1500, certainly within reach of many departments but probably too much for some faculty and most graduate students. Reading oneself into content analysis and through the *Leximancer* manual and training videos is easily done over the course of an academic term. But the real rewards come gradually as the scholar adapts these tools to his/her research agenda and begins to see data everywhere.

REFERENCES

Beeferman, D., Berger, A., & Lafferty, J. (1997). A model of lexical attraction and repulsion. In P. R. Cohen & W. Wahlster (Eds.), Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (pp. 373-380). Madrid: Association for Computational Linguistics.

Clandinin, D. J., & Connelly, F. M. (2000). Narrative inquiry: Experience and story in qualitative research. San Francisco, CA: Jossey-Bass Publishers.

Courtial, J. P. (1989). Qualitative models, quantitative tools and network analysis. Scientometrics, 15(5/6), 527–534.

Grimbeek, P., Bryer, F.**,** Davies, M., & Bartlett, B. (2005). Themes and patterns in 3 years of abstracts from the "International Conference on Cognition, Language, and Special Education Research: identified by Leximancer analysis. In B. Bartlett, F. Bryer, and D. Roebuck (Eds.), **Stimulating the action as participants in participatory research**, 2, 101-113. Brisbane, Australia: Griffith University, School of Cognition, Language, and Special Education.

Indulska, Marta and Recker, Jan C. (2008) Design Science in IS Research: A Literature Analysis. In Gregor, Shirely and Ho, Susanna, Eds. Proceedings 4th Biennial ANU Workshop on Information Systems Foundations, Canberra, Australia.

Lee, B., & Jeong, Y.-I. (2008). Mapping Korea's national R&D domain of robot technology by using co-word analysis. Scientometrics, 77(1), 3–19.

Leximancer Manual v. 3.5 (2010). Downloaded from https://www.leximancer.com/

Leydesdorff, L., & Hellsten, L. (2006). Measuring the meaning of words in contexts: An automated analysis of controversies about 'Monarch butterflies', 'Frankenfoods', and 'stem cells'. Scientometrics, 67(2), 231–258.

Liesch, P., Håkanson, L., McGaughey, S., Middleton, S., and Cretchley, J. 2011. The evolution of the international business field: a scientometric investigation of articles published in its premier journal. Scientometrics, 88 (1), 17-42.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know. verbal reports on mental processes. Psychological Review, 84 (3), 231-259.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning. Urbana, IL: University of Illinois Press.

Sadiq, S., Yeganeh, N. and Indulska, M. (2011). 20 years of data quality research: Themes, trends and synergies. In: Heng Tao Shen and Yanchun Zhang, Conferences in Research and Practice in Information Technology. Proceedings of: The 22nd Australasian Database Conference (ADC 2011). Australasian Database Conference [ADC], Perth, WA, Australia, (1-10). 17-20 January 2011.

Smith, A. and Humphreys, M. (2006). Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. Behavior Research Methods, 38(2), 262-279.

Sowa, J. F. (2000). Knowledge representation: Logical, philosophical, and computational functions. Pacific Grove, CA: Brooks Cole.

Wikipedia (2011). Bayes Factor. Downloaded from http://en.wikipedia.org/wiki/Bayes_factor

Zimitat, C. 2006. A lexical analysis of 1995, 2000 and 2005 ascilite conference papers. In L. Markauskaite et al. (Eds.), Proceedings of the 23rd annual conference of the Australasian Society for Computers in Learning in Tertiary Education: Who's learning? Whose technology? Sydney 3-6 December 2006, 947-951.