Content analyses With Leximancer: Case of ISTE 2011 Conference Abstracts

David A. Thomas
Department of Science, Mathematics and Technology Education, University of Pretoria
david.cynthia.thomas@gmail.com

Abstract
More and more, scholars are turning to automated content analysis technologies to achieve what they do not have time to do themselves, characterize the global features of a large and growing corpus of professional literature and to identify relationships between particular concepts and themes. This paper begins with a brief discussion of content analysis as a research methodology. The utility of this methodology is then illustrated using Leximancer (2010), an automated content analysis technology, to identify major concepts and themes in the 87 abstracts accepted for the 2011 ISTE Conference. Against this conceptual and technical background, the value of content analysis is discussed in a variety of academic contexts, including individual scholarship, departmental and interdisciplinary research groups, and graduate student advising.

1. INTRODUCTION
Digital discourse (e.g., cell phones, email, threaded discussions, synchronous chats, and webinars) and documents (e.g., journal articles, conference proceedings papers, eBooks, research databases, and university-based student and administrative records) have all but replaced older communication technologies (e.g., analogue phones, snail mail, and hardcopy print) in millions of workplaces and homes. Stored in many formats, digital data are easily transmitted, transformed, edited, and re-purposed. Inexpensive word processing, data management, data analysis, presentation, publication, and social networking technologies have fuelled an exponential growth in digital discourse and publication. For instance, Worldometers (2011) estimates that in 2011 two billion internet users sent 300 million emails per day. In the same year, more than one million books were published worldwide.

To cope with this information explosion, scholars in many knowledge domains rely on sophisticated information technologies to search for and retrieve records and publications pertinent to their research interests. But what is a scholar to do when a search identifies hundreds of documents, any of which might be vital or irrelevant to his/her work? The problem is further complicated by the unstructured nature of most documents which, unlike databases, are difficult to search systematically. More and more, scholars are turning to automated content analysis technologies to achieve what they do not have time to do themselves, characterize the global features of a large corpus of work and identify relationships between particular concepts, themes, and methodologies.

This paper begins with a brief discussion of content analysis as a research methodology. The utility of this methodology is then illustrated using Leximancer (2010), an automated content analysis technology. Against this conceptual and technical background, the value of content analysis is discussed in a variety of academic contexts, including individual scholarship, departmental and interdisciplinary research groups, and graduate student advising.

2. CONTENT ANALYSIS
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 content 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; Stubbs, 1996). 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 “content” and “analysis” frequently occur within the same sentence.

3. INVESTIGATING PROFESSIONAL KNOWLEDGE BASES

Leximancer has been used in a number of studies to analyse 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 study sought answers to several questions, including: What percentage of papers at IS conferences pertain to design science research?; Is the focus on design science in IS publication outlets increasing over recent years?; What are the main thematic foci of IS design science papers?; and, Is design science in IS concentrated within schools in specific geographical areas? 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.

Finally, 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 analyse 87 abstracts from the 2011 International Conference on Mathematics, Science, and Technology Education. Prior to undertaking a content analysis of the abstracts, the following research questions were asked.

- Which concepts occur with the greatest frequency? Which concepts are closely associated with one another?
- What are the emergent themes?
- To what extend do the abstracts share common concepts and themes?

4. 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 analysed using Leximancer, provided that the computer on which the software is running is adequate to the task. Once the analysis is complete, three options are available: Concept Map, Insight Dashboard, and Data Exports.

![Figure 1 Leximancer Control Panel](image)
Figures 2 shows a network-like sets of unlabelled nodes and connecting line segments. In this figure, the nodes near the center of the figure correspond to discovered concepts. Further from the center are other nodes associated with the 87 abstracts, coded by first author. This presentation separates the papers themselves from the concepts they contain.

Figure 2 Concept and Paper Nodes

In Figure 3, the concepts discovered by Leximancer are displayed adjacent to their respective nodes. Note that the concepts teachers, school, and learning appear as large dots near the center of the map and the concepts group and practical appear as smaller dots on the periphery. In each concept map,

- 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 Tables 1 and 2. In Table 1, the concepts are listed in ranked order based on the number of times they occur 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. The ten most frequent concepts encountered in this study were teachers, study, students, used, teaching, school, learning, science, learners, and education. In that list, we see that the concepts teachers and study occur, respectively, 183 and 142 times. When a concept in the ranked list is clicked, such as teachers, another screen is displayed (see Table 2) showing the co-occurrence of the concept with other discovered concepts, such as professional. In that line of the table, we see that technology co-occurred with professional 21 times. The Likelihood score of 78%, computed as (# co-occurrences of teachers and professional)/(# occurrences of professional), may be interpreted as saying that the concept teachers co-occurs with the concept professional 40% of the time. Clicking on the magnifying glass icon to the left of the professional concept opens a window in which each of those 21 co-occurrences is displayed as seen in the original abstracts.
The Thesaurus tab shows which words are associated with any given concept. For instance, in Table 3, 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 Table 4). The “Iterations: 1” notation indicates that 1 iteration of learning was used to generate a stable set of concept definitions.

Table 3 Thesaurus Word Definitions

Table 4 Thesaurus Source Texts
Themes are clusters of concepts. Like concepts, themes are heat-mapped, meaning hot colours (red, orange) denote the most relevant themes, and cool colours (blue, green), denote the least relevant. In this study, Leximancer discovered 4 major themes seen in tabular format in Table 5, teachers, study, learners, and practical. Theme connectivity is the summed co-occurrence counts of each concept within the theme with all available concepts.

Table 5 Discovered Themes in Tabular Format

<table>
<thead>
<tr>
<th>Theme</th>
<th>Connectivity</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>teachers</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>study</td>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>learners</td>
<td>25%</td>
<td></td>
</tr>
<tr>
<td>practical</td>
<td>12%</td>
<td></td>
</tr>
</tbody>
</table>

The map seen in Figure 4 shows the major themes in graphical format. The % Theme Size slider under the Concept Map may be used to adjust the number and relative sizes (i.e., generality) of the themes displayed, reducing the overall complexity of the display when all discovered themes are displayed (See Figure 5).
Figure 6 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.
In Figure 7, the papers are positioned on the periphery of the network of nodes and labelled using the first letter of their authors’ last names.

Figure 7 Papers by Author

Overlaying the paper/author data on the concept/theme representations provides insights into where the themes originate (See Figure 8). In this representation, authors appear closest to the concepts and themes found in their abstracts.

Figure 8 Concepts, Themes, and Authors

In Figure 8 different authors are associated with different themes. Interpreted in terms of the forest metaphor, think of the concepts as different species of trees (e.g., oak, maple, pine), the
themes as different landscapes where they grow (e.g., meadow, hillside, wetland), and the papers/authors as different species of birds (e.g., robins, starlings, sparrows) that nest in the trees. Taken one step further, our metaphor says that while each species of bird (i.e., each author) prefers to nest in a particular species of tree (i.e., pursue a particular research interest) growing in a particular landscape (i.e., situated in a particular knowledge domain), they often take a broad view of things.

As seen in Figure 6 Emergent Concepts and Themes, the concept map uses the distance between concepts nodes as a metaphor for the “distance” (i.e., similarity/differences) between the concepts themselves. Recalling that each concept is defined in the Leximancer thesaurus by a list of weighted words, two concepts may be contrasted by comparing their respective word lists. The Pathways tab (See Figure 9) provides a graphical interpretation of such comparisons. Clicking two concepts (e.g., teaching and learners) illustrates the pathway between them, along with example text. The relationships between concepts are best thought of as correlations, though the text segments describing the relationship may define a direction for cause. Clicking on the more... link in the text window in Table 7 reveals the occasions that these words occur together and the measure of association, 0.74.

Figure 9 Concept Pathway Graphic
As seen in Figure 10 and Table 7, a similar procedure can be used to map pathways between entire documents. In this case, the connection is stronger, with a measure of association of 0.87. These papers (and perhaps the authors) share common interests.
Leximancer’s Insight Dashboard provides a number of other outputs, both tabular and graphical in nature. These outputs make reference to measures identified as frequency, strength, and prominence. Frequency is coded as a conditional probability the chance that a given attribute is coded in a text excerpt. This corresponds to frequency of mention in the data, and is affected by the distribution of comments across the Categories. The strength score is the reciprocal conditional probability. Given that a particular attribute is present in a section of text, it gives the probability that the text comes from that category. Strong concepts distinguish the category from others, whether or not the attribute is mentioned frequently. Finally, the prominence score is computed as the product of the strength and frequency scores. Prominent concepts are not incidental, they are important ideas. Table 8 shows these measures for three papers.

Table 8 Prominence of Concepts for Selected Abstracts

<table>
<thead>
<tr>
<th>Category: FILE_a_a</th>
<th>Category: FILE_a_e_r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept</td>
<td>Concept</td>
</tr>
<tr>
<td>Rel Freq (%)</td>
<td>Rel Freq (%)</td>
</tr>
<tr>
<td>Strength (%)</td>
<td>Strength (%)</td>
</tr>
<tr>
<td>Prominence</td>
<td>Prominence</td>
</tr>
<tr>
<td>technology</td>
<td>technology</td>
</tr>
<tr>
<td>38</td>
<td>40</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>science</td>
<td>science</td>
</tr>
<tr>
<td>50</td>
<td>&lt; 1</td>
</tr>
<tr>
<td>4</td>
<td>&lt; 1</td>
</tr>
<tr>
<td>mathematics</td>
<td>mathematics</td>
</tr>
<tr>
<td>38</td>
<td>&lt; 1</td>
</tr>
<tr>
<td>3</td>
<td>&lt; 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category: FILE_b_o</th>
<th>Category: FILE_bc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept</td>
<td>Concept</td>
</tr>
<tr>
<td>Rel Freq (%)</td>
<td>Rel Freq (%)</td>
</tr>
<tr>
<td>Strength (%)</td>
<td>Strength (%)</td>
</tr>
<tr>
<td>Prominence</td>
<td>Prominence</td>
</tr>
<tr>
<td>technology</td>
<td>technology</td>
</tr>
<tr>
<td>22</td>
<td>&lt; 1</td>
</tr>
<tr>
<td>6</td>
<td>&lt; 1</td>
</tr>
<tr>
<td>mathematics</td>
<td>mathematics</td>
</tr>
<tr>
<td>33</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 9 shows the relative prominence of pairs of co-occurring concepts for two of these papers. Not surprisingly for an ISTE Conference, technology and mathematics and technology and science are strongly prominent in both.

Table 9 Prominence of Co-occurring Concepts in Selected Papers
ANSWERING THE RESEARCH QUESTIONS

The questions asked relative to the 2011 ISTE abstracts and their respective answers are as follows:

- **Which concepts occur with the greatest frequency?**
  The ten most frequent concepts encountered in this study were teachers, study, students, used, teaching, school, learning, science, learners, and education.

- **Are closely associated with one another?**
  The Insight Dashboard report, which is too massive to address in this paper, shows a consistent pattern of association between concepts related to teaching, technology, science, and mathematics.

- **What are the emergent themes?**
  The emergent major themes are teachers, study, learners, and practical. Again, this is hardly a surprising list for abstracts at an ISTE Conference.

- **To what extent do the abstracts share common concepts and themes?**
  From Figures 7 and 8 suggests a complex web of relationships with different abstracts reflecting different priorities relative to the major themes and concepts. This is to be expected. The real interest here is which particular authors share particular interests. Armed with that insight, these scholars may find motivation for future collaborations.

6. IMPLICATIONS FOR INDIVIDUAL SCHOLARSHIP, COLLABORATION, AND GRADUATE STUDENT ADVISING

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.

7. REFERENCES
Leximancer Manual v. 3.5 (2010). Downloaded from https://www.leximancer.com/
Worldometers – Real Time World Statistics (2011). Downloaded from worldometers.info