Available online at www.sciencedirect.com
ScienceDirect
Int. J. Human-Computer Studies 67 (2009) 424–436
International Journal of Human-Computer Studies
www.elsevier.com/locate/ijhcs

Use of an automatic content analysis tool: A technique for seeing both local and global scope

Paul Stockwell[a,*], Robert M. Colomb[a], Andrew E. Smith[b], Janet Wiles[a]

[a]School of Information Technology and Electrical Engineering, The University of Queensland, St. Lucia, QLD 4072, Australia
[b]Institute for Social Science Research, The University of Queensland, St. Lucia, QLD 4072, Australia

Received 29 August 2007; received in revised form 10 November 2008; accepted 7 December 2008
Communicated by N. Aussenac-Gilles
Available online 13 December 2008

Abstract

This paper examines what can be learned about bodies of literature using a concept mapping tool, Leximancer. Statistical content analysis and concept mapping were used to analyse bodies of literature from different domains in three case studies. In the first case study, concept maps were generated and analysed for two closely related document sets—a thesis on language games and the background literature for the thesis. The aim for the case study was to show how concept maps might be used to analyse related document collections for coverage. The two maps overlapped on the concept of "language"; however, there was a stronger focus in the thesis on "simulations" and "agents." Other concepts were not as strong in the thesis map as expected. The study showed how concept maps can help to establish the coverage of the background literature in a thesis. In the second case study, three sets of documents from the domain of conceptual and spatial navigation were collected, each discussing a separate topic: navigational strategies, the brain's role in navigation, and concept mapping. The aim was to explore emergent patterns in a set of related concept maps that may not be apparent from reading the literature alone. Separate concept maps were generated for each topic and also for the combined set of literature. It was expected that each of the topics would be situated in different parts of the combined map, with the concept of "navigation" central to the map. Instead, the concept of "spatial" was centrally situated and the areas of the map for the brain and for navigational strategies overlaid the same region. The unexpected structure provided a new perspective on the coverage of the documents. In the third and final case study, a set of documents on sponges—a domain unfamiliar to the reader—was collected from the Internet and then analysed with a concept map. The aim of this case study was to present how a concept map could aid in quickly understanding a new, technically intensive domain. Using the concept map to identify significant concepts and the Internet to look for their definitions, a basic understanding of key terms in the domain was obtained relatively quickly. It was concluded that using concept maps is effective for identifying trends within documents and document collections, for performing differential analysis on documents, and as an aid for rapidly gaining an understanding in a new domain by exploring the local detail within the global scope of the textual corpus.
© 2008 Elsevier Ltd. All rights reserved.

Keywords: Leximancer; Concept; Mapping; Local; Global; Scope; Content; Analysis

1. Introduction

The 21st century is the information age. Not only are vast quantities of data relevant, but digital formats making documents readily accessible. The rapid growth in available literature is not matched by a comparable increase in available time or cognition. It is not necessary to read all the papers in a body of literature to come to some understanding of it; a variety of rudimentary methods can be employed, such as perusing reviews, or even skim reading or more sophisticated techniques including automatic text summarisation. An alternative to these methods is content analysis, which is based on statistical processing of text corpora (Weber, 1990).

Content analysis extracts the most commonly occurring terms in a body of literature and tabulates the frequency of their co-occurrence into a matrix (Berelson, 1951). As

*Corresponding author. Tel.: +61 733 651 654; fax: +61 733 654 999.
E-mail addresses: stockwell@itee.uq.edu.au (P. Stockwell), colomb@itee.uq.edu.au (R.M. Colomb), a.smith7@uq.edu.au (A.E. Smith), j.wiles@itee.uq.edu.au (J. Wiles).

1071-5819/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijhcs.2008.12.001

technology improves, the ability to perform more detailed analysis on the resulting term co-occurrence using visual tools has become available. With change in technology has come change in usage. Content analysis has been applied in a variety of ways, and several techniques have been developed. Hyperspace analogue to language (HAL) (Burgess and Lund, 1997), Latent semantic analysis (LSA) (Landauer et al., 1998) and Leximancer (Smith, 2000) are common algorithms that take a corpus of text and extracts the key terms to create a co-occurrence matrix of their relationships.

The concepts within a domain may be represented visually. Concept maps are a standard method of graphically displaying the relationships between concepts in a domain (Novak and Gowan, 1984). Concept maps may be derived manually as a mind map, or algorithmically from a corpus of text. A mind map is an individual's own internal representation of a domain and can be useful in structuring the nature of the domain. Algorithmic concept maps have been created using techniques such as correspondence analysis (CA) (Benzecri, 1992) or self-organising concept maps (SOCOMs) (Hagiwara, 1995), and available tools include Tetralogie (Mothe et al., 2006), Terminoweb (Barrière and Agbago, 2006), and Leximancer (Smith, 2000).

Text mining tools and concept maps are a well-established method for extracting key concepts from a textual corpus and displaying them in a graphical representation for further analysis. What has not yet been examined is how concept maps may be used to view the global scope of the domain, yet facilitate drilling down to the detail contained within the corpus. The reviewing of a concept within its global context is a real-world application of modern information systems.

One method that may be employed to address global and local scope is to use sets of related concept maps and analyse where they overlap, and how they cover their global domain. Another method is to use a concept map to try to quickly come to a basic understanding of a domain with no a priori knowledge of that domain. The question is, can visual representations such as concept maps aid in rapidly coming to an understanding of the global context within a domain?

The purpose of this paper is to show real-world applications of modern information systems and how they can be used to explore a domain based on a textual corpus from both the high-level view of a domain to specific details of a region contained within it. A series of case studies using concept maps with different real-world conditions can demonstrate the advantage of these tools for maintaining a view of local and global scope-concept maps reveal the context of terms used within a corpus of text and how they fit within their wider domains.

This paper describes how visual representations of concept space can be used to review the global scope of the underlying domain and allow searching of local scope around specific concepts. A series of case studies were conducted using Leximancer (Smith, 2000), a content analysis and concept mapping tool, for comparing multiple document sets and their coverage of smaller, local regions of the domain in the context of the wider domain. For an overview of Leximancer see Section 2.2, and for a full description see Smith and Humphreys (2006).

Leximancer was chosen for a number of reasons: it is implemented as a commercial-quality program, is easy to use and has been evaluated for stability, reproducibility and correlative validity (Smith and Humphreys, 2006). The specific aims for this study are to (a) describe the use of Leximancer to analyse the coverage of related document collections, (b) explore emergent patterns in a set of related concept maps that may not have been apparent from reading the literature alone, and (c) to present concept maps as a tool for rapidly acquiring a basic vocabulary in an unfamiliar domain. Each of these aims motivated a case study on how Leximancer can be used to highlight features of bodies of literature. In the first case study, we examined a thesis and its background literature, investigated the overlap between the two related document sets, and identified where the thesis extended the focus of the background literature (see Section 3). In the second case study, we examined a body of literature divided into three topics related by a central theme, and investigated how the coverage for each of the three sub-topics related to the combined literature set (see Section 4). In the third case study, we show how a Leximancer map helped in successfully crossing disciplinary boundaries (see Section 5). The paper concludes with a summary of results and general discussion.

2. Background literature

2.1. Content analysis and visualisation tools

The creation, visualization, and analysis of conceptual space may be examined in separate components. A corpus of text can be analysed using a form of natural language processing such as content analysis to identify the key terms present. The relationships between the concepts in a domain may be represented graphically. Concept maps are common tools for conceptualising the structure of a domain. However, a concept map may be produced by an individual as a representation of their understanding of the domain, or algorithmically using a co-occurrence matrix derived from a corpus of text. Lastly, a concept map—whether derived manually or algorithmically—may be analysed for gaining insight into the domain represented by it.

Content analysis is the tabulation of the most commonly occurring concepts or terms in a body of literature, typically with the frequency of occurrence attached (Berelson, 1951). The series of books by John Naisbitt and collaborators beginning with *Megatrends 2000* (Naisbitt and Aburdene, 1990) are a well-known example. Their premise is that the number of column-inches available for

news in the USA is generally constant from year to year. An article about a given concept will take up a given amount of space. So as concepts gain prominence, other concepts must lose. The most prominent concepts in a period therefore give a picture of the most significant concerns of the mass media in that period, and changes in significant concerns can be tracked from period to period. Content analysis of all the newspapers published in the USA over a period of time can identify the topical concepts, which can then be used without the necessity of reading the articles.

Content analysis was practiced historically by trained human analysts attaching standardised concept codes to blocks of text. More recently, both experts and untrained individuals have benefited from the availability of inexpensive computing facilities and text in machine-readable form. Relational content analysis (Weber, 1990) tabulates not only the frequency of concepts in the body of text, but also the co-occurrence of concepts in small fragments of text. A co-occurrence matrix is derived, containing the number of co-occurrences between key concepts. The resulting tabulation shows not only the most common concepts but a measure of relationship among concepts derived from co-occurrence.

A number of related methods have been developed, including hyperspace analogue to language (Burgess and Lund, 1997), latent semantic analysis (Landauer et al., 1998), and Leximancer (Smith and Humphreys, 2006). HAL is very close to the raw co-occurrence matrix. The relationships found are symmetric so do not take into account the relative frequency of terms. Further, if words do not co-occur, then HAL assumes they are not related. In particular, synonyms rarely co-occur. LSA is based on principal components analysis. The co-occurrence matrix is processed to extract its eigenvalues and eigenvectors in a process similar to factor analysis. The variance of the eigenvalues is independent of their size, so their statistical reliability decreases with their size. LSA chooses the eigenvectors corresponding to the eigenvalues large enough to be statistically reliable, and then computes an approximation to the co-occurrence matrix from the chosen eigenvectors and eigenvalues. The resulting approximation can show relationships among terms where none existed in the original co-occurrence matrix (Landauer et al., 1998). The explanation is that significant relationships are obscured by noise. Like HAL, LSA is symmetric and therefore cannot utilise relative frequency with co-occurrence. Paradoxically, LSA has better agreement with human interpretation when less data are available to it. The variance of eigenvalues decreases as sample size increases, so with a larger body of text smaller eigenvalues are reliable. The more text analysed, the smaller the effects that can be seen unless the smaller statistically reliable eigenvalues are arbitrarily discarded. Leximancer, a content analysis and concept mapping tool, is an asymmetric method, so takes frequency of occurrence into account, and is capable of extracting relationships between words such as synonyms that are semantically related but rarely co-occur. It is also a direct method, so all of a body of text contributes to the result.

Visualising a domain and the relationships between the key concepts graphically can aid in learning and aiding in comprehension. Concept maps (Novak and Gowan, 1984) are one method for visualising these relationships spatially. Concepts are drawn as points or circles in a two-dimensional space, and connections between concepts are represented by lines drawn between them. Concept maps can be created manually, or they can also be generated algorithmically from a corpus of text. Case studies on the use of concept maps as a learning tool have shown that they are effective in refining ideas and how they fit together (Novak, 1990). Bibliometric maps show clusters of papers and the relationships between them (Buter and Noyons, 2002). ''Non-content-bearing'' phrases that are common to many fields are discarded and the distribution into clusters of ''content-bearing'' phrases can provides an overview of the authors for each scientific field.

Once the key concepts have been extracted from a corpus of text in natural language, their co-occurrence relationships can then be visualised and interpreted for further study. Several tools and algorithms can be used to derive a concept map or visual representation of key terms from a corpus of text. Correspondence analysis (Benzecri, 1992) has been used for on a wide range of visualisations where there is a requirement for dimensionality reduction. A co-occurrence matrix can be presented as a two-dimensional plot of the relationships of the key terms with correspondence analysis. Self-organising concept maps use neural networks to arrange terms in a map (Hagiwara, 1995). Two different types of SOCOMs were tested; metric using metric data such as similarity, and non-metric based on the rank order of similarity among items. It was shown in simulation that the SOCOMs were able to arrange terms in a map space and did not require all detailed metric information. Other techniques that have been utilised for visualising concept space are described in Chen (2002).

There are several tools available that extract data algorithmically from a corpus of text and display their relationships graphically for further analysis. Tétralogie (Mothe et al., 2006) identifies trends in scientific communities from large document collections by geographical location using a set of agents to implement their mining and visualisation tools. Classification is achieved using agglomerative hierarchical clustering and classification by partition. Graphical representations of results are displayed to the user as networks, histograms, and geographical maps. Other tools perform clustering of documents from the Internet based on the conceptual relationships contained within them rather than geographic origin. CruxLux (http://www.cruxlux.com) searches the Internet for text documents such as web pages and weblogs that match given search criteria, and performs a graphical clustering of the documents based on similarity of content. TerminoWeb (Barrière and Agbago, 2006) is designed to

perform thematic searches of documents on the Internet. It uses third-party Web search engines to build a corpus of documents that are measured on their textual flow, where a natural language-like text structure is preferred. Content exploration is achieved using knowledge-rich contexts (Barrière, 2004). "Knowledge-rich" documents with a high density of knowledge patterns contain many domain specific terms. Term frequency and a reference general language corpus are then used for term extraction. Rather than clustering on documents, there are also applications that create graphical representations of the relationships between concepts contained within documents or document collections. Mindsystems ThemeReader (http://www.mindsystems.com.au) can scan a provided document and represent the key terms in a hierarchical tree or map layout. Key terms are identified by occurrence frequency. Documents are clustered based on their similarity of key terms. Leximancer (Smith, 2000) uses content analysis to extract the co-occurrence matrix from a collection of documents and displays the relationships between terms graphically in a two-dimensional map. Further analysis of total occurrences and co-occurrences of key terms and drill-down to the original text may be performed by the user.

The problem of extracting information from large or complex corpora of text has been studied extensively (Salton, 1989; Weber, 1990). Data may be mined for key terms or bibliometric collaboration, and their relationships analysed. Concept maps and other visual representations of a domain are useful when manually generated such as mind mapping an individual's internal understanding, or derived algorithmically using some form of content analysis or natural language processing to extract the key terms from a corpus of text. These visual representations have been analysed using a variety of techniques including bibliometrically, geographically and term co-occurrence. Regardless of technique, the goal is to reach a greater understanding of the domain being represented.

2.2. Leximancer

Leximancer is a semi-automatic content analysis tool that can be used to analyse a document or collection of documents (Smith, 2000). It is capable of identifying key terms using word frequency and co-occurrence usage. Relational content analysis (Weber, 1990) is used on the data consisting of episodic co-occurrence records. The classifier is based on a generalisation of a Bayesian classifier (Yarowsky, 1995). Stemming is not performed automatically; however, the user may choose to combine word stems into a single term manually. Terms that are less frequent yet used consistently with a more common term are grouped as term classifiers in a thesaurus and are referred to as *concepts*. Leximancer can then show the relationships between concepts graphically in a concept map (see Fig. 1). These relationships can be asymmetric: the co-occurrence between two concepts may be relatively stronger in one direction than the other, depending on the total occurrence for each concept.

The Leximancer mapping subsystem works in two stages, characterised as semantic extraction followed by relational extraction (Smith and Humphreys, 2006). For each stage, the data consist of actual episodic co-occurrence records. In the first stage, called semantic extraction, a document-by-term network is first computed using small text segments which consist of windows of 1–3

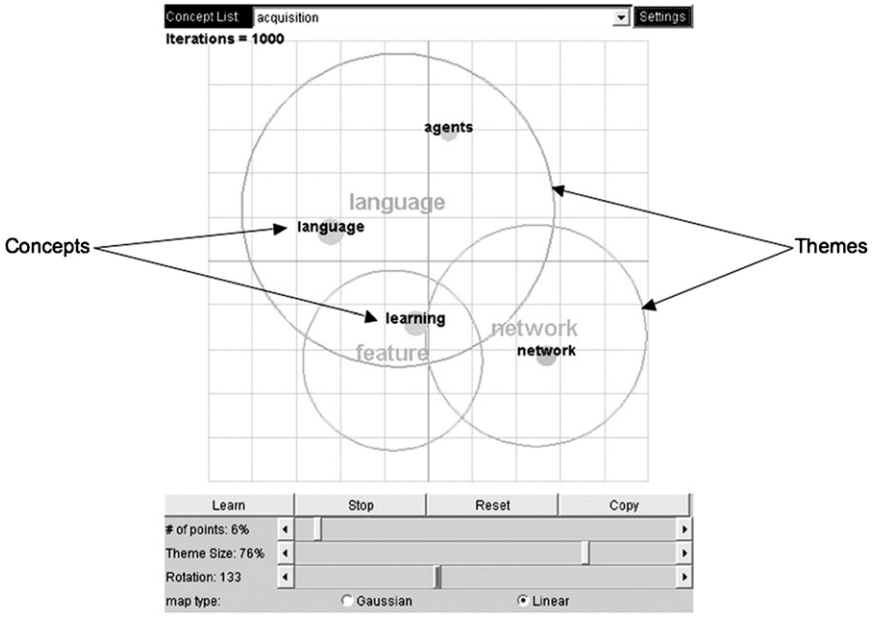

Fig. 1. Sample Leximancer concept map. The size of each concept point is determined by the summed co-occurrence counts with all other concepts. The number of points that appear on the map can be controlled using the "# of points" slider. Themes are groupings of concepts and are controlled by the "Theme Size" slider.

sentences. These text segments may cross paragraph boundaries or can be restricted to within paragraphs (default setting). The semantic representation needs to be capable of classifying small segments of text with limited available evidence, which led to the selection of a naïve Bayesian co-occurrence metric (Salton, 1989) for co-occurrence relevance. This metric is known to perform well as a text classifier (Dumais et al., 1998). The co-occurrence relevance function $g(x)$ uses Bayes' theorem:

$$\log g(x) = \log \frac{Pr(x|rel)}{Pr(x|nonrel)} + \log \frac{Pr(rel)}{Pr(nonrel)'} \quad (1)$$

for each record x where $Pr(rel)$ and $Pr(nonrel)$ are a priori probabilities of the relevance (*rel*) and non-relevance (*nonrel*) for any record (Salton, 1989). The semantic extraction stage utilises a concept bootstrapping algorithm developed from a word sense disambiguation algorithm to identify families of weighted terms which tend to appear together (Yarowsky, 1995). These families of terms form predictive classifiers which are used to tag text segments with likely concepts. Eq. (2) is derived from Eq. (1) under the naïve Bayesian approximation. The term-relevance weight, tr_j, for each term x_j is then given by

$$tr_j = \log \frac{Pr(x_j = d_j|rel)Pr(x_j = 0|nonrel)}{Pr(x_j = d_j|nonrel)Pr(x_j = 0|rel)} \quad (2)$$

where d_j is the term weight for term x_j, and is determined by the occurrence probabilities of the term in the relevant and non-relevant items (Salton, 1989).

The second stage, called relational extraction, begins by classification of text segments using the learned concept classifiers. From this information, a matrix of concept co-occurrence counts is calculated. On the diagonal of this matrix is the occurrence count for the relevant concept. Each column of the co-occurrence matrix is divided by its diagonal, making a cell the probability of occurrence of the row term given the occurrence of the column term. The conditional probability $Pr(x)$ of occurrence of concept x_i given concept x_j is approximated by

$$Pr(x) = \frac{c_{ij}}{f_i} \quad (3)$$

where c_{ij} is the co-occurrence frequency of concepts x_i and x_j, and f_i the frequency occurrence count for concept x_i. The resulting concepts and their conditional probabilities are then used as initial conditions for particles and forces (respectively) in a highly dissipative iterative numerical model. The Leximancer concept mapping algorithm is based on a variant of a spring-force model for the many-body problem (Chalmers and Chitson, 1992). The calculation produces a set of positions coerced into a planar geometry. Distance between concepts in this abstract space is taken as a measure of relatedness: the smaller the distance the closer the relationship. The context in which concepts are used in the original text determines the position of concept points. When a concept point is clicked, connecting lines represent its co-occurrence.

Because the number of concepts is typically large, they are grouped by proximity into clusters, which are called *themes*. Themes are given default names from the most central concept. Leximancer has been used on many large bodies of text, and is computationally tractable. For further details of the algorithm see Smith and Humphreys (2006).

Different levels of details can be displayed on a Leximancer map: varying the number of concept points changes the number of visible concepts, but has no effect on the themes. Changing the theme size by moving the slider alters the layout of the themes (radii and centroids), from one theme per concept when set to 1% (or none at 0%), to one single theme encapsulating every concept when set to the maximum value of 99%. Further analysis can be used to quantify the concept map based on a ranked concept list, and the original text corpus can be browsed directly via the co-occurrence matrix.

Initial exploration of the concept map may be achieved in several different ways in Leximancer. One method is to hide all of the theme circles and show only the most significant concepts initially, and then slide the point slider toward 100% to reveal more concepts on the map. Another method is to start with the theme slider at or near 99%, showing a small number of themes. Because the themes are based on centroids and radii, closely positioned significant concepts will coalesce into a single theme, while other concepts that are not as significant but are more distantly positioned will receive their own theme. Moving the theme slider toward 0% decreases the size of the radii, thus producing more numerous, more specific themes. Regardless of which method is used, the goal is to start at the high-level view of the represented domain, then explore with increasing detail until the desired level has been reached. The user may then investigate the relationships between concepts, the text segments or the terms in the document collection that comprise the thesaurus.

Although using the default settings can result in a useful map, some elements of the analysis are semi-automatic and may be adjusted by the user to add value. The user may choose to increase or decrease the number of terms identified in the corpus, to merge word stems, to remove terms from being included into concepts, or not to use the thesaurus. The process is iterative; as the map is updated to reflect the analysis choices, the user may further refine the map to suit their requirements.

Two types of maps can be generated: linear or Gaussian. Linear maps often produce a more clustered structure and are useful for highly connected networks as they are less influenced by individual activity. Gaussian maps normalise the weights of attraction between concepts and tend to be more uniformly distributed than linear maps and can obscure the strength of relationships between concepts. Linear maps were therefore used with Leximancer version 2.21 for all of the case studies.

3. Case study 1: a thesis and its background literature

The aim for the first case study was to examine how a concept map can highlight where a thesis overlaps its background literature, and where the thesis extends the literature on which it was based. Concept maps were created and analysed for an undergraduate thesis "Language Games and Generalisation Grounded in Autonomous Agents" (Stockwell, 2005), and its background literature. The studies for the thesis were conducted as part of the RatChat project, an ongoing project investigating simulated language learning using mobile agents (Schulz et al., 2008, 2006) that extends work on simultaneous localising and mapping based on the rat hippocampus (Milford et al., 2004). Stockwell (2005) was selected due to the high level of familiarity with the topic and literature, and its coverage across interdisciplinary boundaries of highly cited papers.

The primary expectation for this case study was that there would be a high level of overlapping coverage of concepts in both the thesis and its background literature. It was also expected that the key concepts in the thesis would be "language", "generalisation", and "environment", and that these concepts would be present on the thesis map.

3.1. Method

Concept maps were generated in Leximancer for both the thesis and for 25 electronically available papers from the background literature. For both maps, spurious concepts from the bibliography or the text were removed over a number of iterations; in each map the number was less than five. Synonyms and word stems were manually merged (e.g. 15 concept groups were formed from 32 concepts for the background literature). The co-occurrence matrices for both maps were then compared statistically.

3.2. Results

3.2.1. Concepts

The thesis map showed that the primary concepts in order of frequency were "language", "simulations", "learnability", "agents", "generalisable", and "listener" (see Fig. 2b). The top five primary concepts were grouped tightly together, with "listener" located separately (see Fig. 2a).

The background literature map had the primary concepts in order of frequency as "language", "learning", "network", "agents", "evolution", and "system" (see Fig. 2d). A uniform distribution of the primary concepts was apparent with no specific clustering (see Fig. 2c).

3.2.2. Themes

For the thesis map, the "language" theme encapsulated all of the primary concepts with the exception of "listener", which was contained in its own theme, which intersected with the "language" theme (see Fig. 2a). The "research" theme intersected both the "language" and "listener" themes and contained the concept of "simulations". The smaller "transmission" theme intersected only with the "language" theme and contained none of the primary concepts. The background literature map was more spread out with "language", "evolution", and "system" in the "language" theme (see Fig. 2c). "Agents" was contained in its own theme intersecting with the "language" theme and also contained the "system" concept. "Learning" was situated in its own theme, which intersected the "agent", "language", and "network" themes. The smaller "fitness" theme intersected only with the "language" theme and contained none of the primary concepts.

3.2.3. Co-occurrence

Clicking on each concept in turn and examining the relative strength of the connecting line between concepts indicated its co-occurrences. The thesis showed a strong co-occurrence between the top five concepts, and all showed a smaller co-occurrence with the concept of "listener". However, a strong asymmetric relationship was apparent between "listener" and "simulations," "language", and "agents". All of the primary concepts co-occurred in the background literature, with "system" showing a strong co-occurrence with all other primary concepts. "Language", "learning", and "network" co-occurred frequently, while "evolution" showed a strong co-occurrence with "learning" and "language" that always occurred symmetrically. "Language" co-occurred frequently only with "learning," which was also asymmetric.

3.3. Discussion

With the goal of examining what can be learned about sets of literature using concept maps, the first case study examined how a concept map can highlight where a thesis overlaps and extends its background literature. Both maps had "language" as the primary concept as represented by the largest concept point. "Learning"/"learnability", and "agents" were also both highly ranked (shown with large concept points in Fig. 2a and d, and in the ranked concept lists, Fig. 2b and c). However, "agents" was not as closely related to "language" in the background literature map and were further apart on the map. The expected primary concepts for the thesis were visible on the map; however, "simulations" was also a strong concept in the ranked concept list. The unexpected appearance of "simulations" as a strong concept was investigated using Leximancer's tools, and had a strong co-occurrence with "language," "learnability", and "generalisable" (co-occurrence rays not shown on figure for clarity). Asymmetrically, "simulations" was the most commonly co-occurring concept for "language" and the second most commonly co-occurring concept for both "learnability" and "generalisable," indicating that it had a strong global significance within the document.

There is positive correlation for the relative strengths in connectivity between the concepts that are common in both

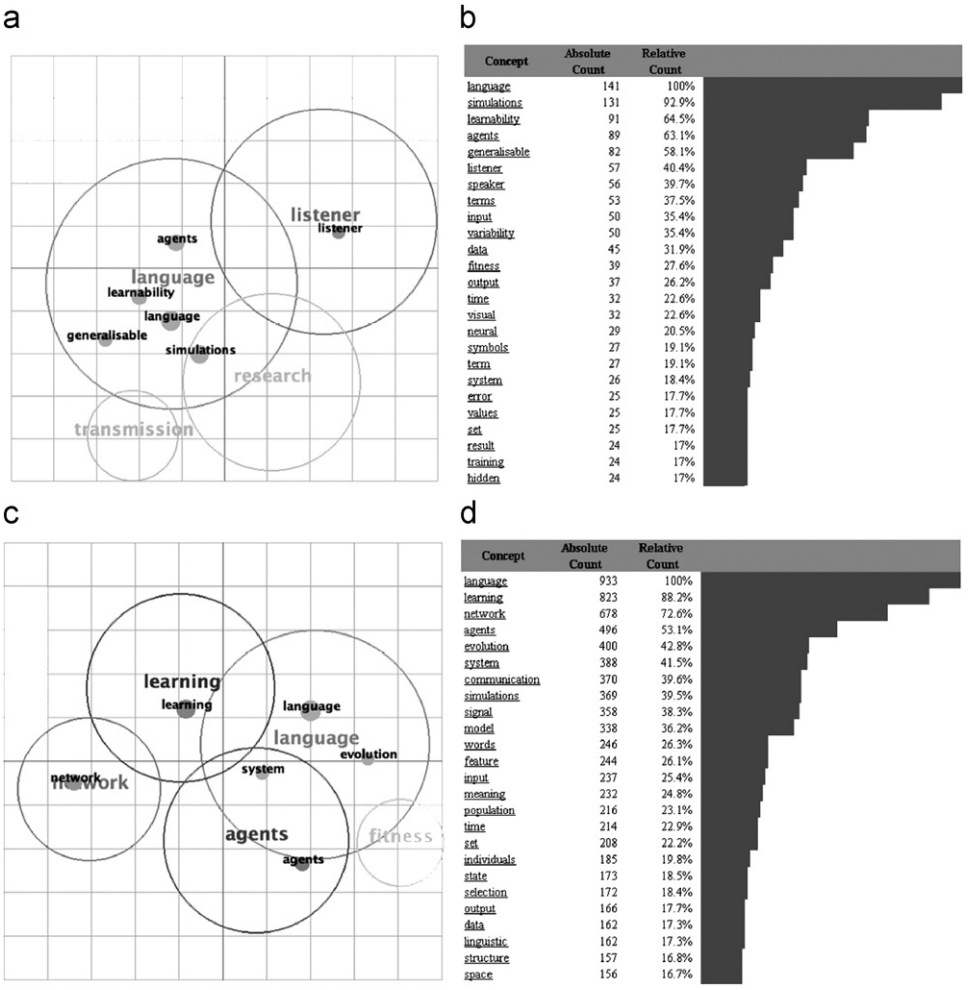

Fig. 2. Leximancer map and concept terms for case study 1: (a) map for the RatChat thesis at 60% theme size. Up to 60 concepts can be shown and only six are shown for clarity. The theme "transmission" contained less significant concepts, not shown; (b) most common terms for the thesis; (c) map for the literature collection at 40% theme size. Up to 75 concepts can be shown and only seven are shown for clarity; and (d) most common terms for the literature collection.

maps (see Fig. 3). There is high connectivity for "language" in both maps, but "simulations" is significantly higher and "agents" is slightly higher in the thesis map than the background literature map, indicating a focus in the thesis on agent-based simulations on language.

The more uniform distribution of concepts in the literature map (Fig. 2c compared with 2a) is likely to be due to the larger set of documents than the 12,000 word thesis. A larger document collection covers a greater scope of material and would tend to be less focussed on a particular set of concepts, although if the documents were all selected on a given topic that theme would have a high level of co-occurrence. This behaviour is evident in Fig. 3: "Language" is the central theme for the document collection, but although "simulations" and "agents" still show a high interconnectivity, in general the relationships are less pronounced than in the thesis.

A thesis is intended to show the coverage of the relevant parts of the literature, and then extend it in a new way. The dual concept maps can show which parts of the overall literature has been successfully covered by the thesis literature review, and also where the thesis emphasises points or extends the topic to new concepts or ideas. By using a pair of concept maps, information covered from the literature review by the thesis can be confirmed, and additionally what is not covered that perhaps should have been can be identified, assuming a broad enough literature set. The local context of a thesis map can be established within the global context of the literature map.

4. Case study 2: a multidisciplinary topic: conceptual and spatial navigation

The first case study investigated the coverage of a document that intersects a document collection. The second case study examined the coverage of a concept map that included all of the documents in a related document collection, divided into sub-domains. The Thinking Systems project, which examines navigation in real and conceptual space, was used as the domain. Three sub-domains: navigation and the brain; conceptual and spatial

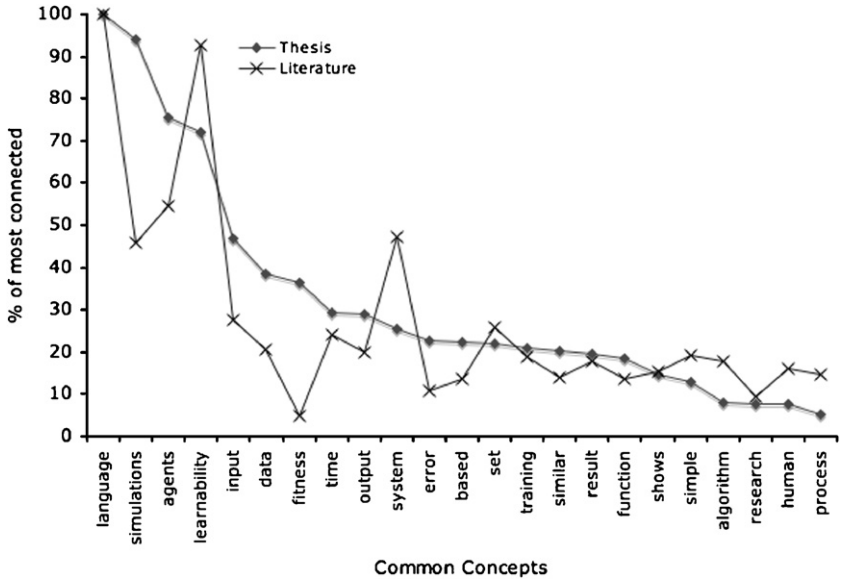

Fig. 3. Correlation of most connected concepts for concepts that appeared in both maps in case study 1. For all concepts that are present in both maps, the number of connections was calculated as a percentage of the most connected concept and is shown on the *y*-axis. The line shown with diamonds represents the thesis map, and the line with crosses the literature map. The set of common concepts is shown in the *x*-axis, in descending order of significance from the thesis map.

mapping; and physical navigational strategies were selected in the context of the wider project.

The aim of this case study was to see if the distinction between three disparate, though potentially related, sub-domains is reflected in the coverage of the map, with a primary expectation that the map would show each of the three sub-domains in different parts of the map. A secondary expectation was that the concept of ''navigation'' would be central to the combined map as the interconnecting theme.

4.1. Method

Combinations of key words central to the Thinking Systems project were used as criteria in Google Scholar (scholar.google.com), for each topic. The highest ranked documents most appropriate to the connecting theme of navigation were selected—five documents for physical navigation techniques using the keywords ''navigation'', ''techniques'', five documents for the brain and navigational tasks with the keywords ''navigation'', ''amygdala'', ''hippocampus'', and ''parietal cortex'', and eight documents for conceptual and spatial mapping using the keywords ''navigation'', ''concept'', ''conceptual'', ''spatial'', and ''map.'' The set of specific search terms for each topic were selected as part of the common terms.

A separate concept map was generated in Leximancer for each sub-domain. As in the first case study, spurious concepts and names from bibliographies were removed and similar concepts such as words and their plurals were merged. A concept map was then generated for the combined literature set and refined by removing bibliography entries and merging related concepts. The primary concepts in each of the sub-domain maps were identified and their relative positions on the combined map analysed.

4.2. Results

4.2.1. Concepts

The primary concepts identified on the physical navigation map in order of frequency of occurrence were ''rats'', ''location'', ''strategy'', ''hippocampus'', and ''navigation'' (see Fig. 4a). For the map on the brain and navigation, the primary concepts in order of frequency were ''hippocampal'', ''rats'', ''animals'', ''stimulation'', and ''studies'' (see Fig. 4b), while for the conceptual and spatial mapping map, the primary concepts in order of frequency were ''symbol'', ''concepts'', ''system'', ''representations'', and ''cognitive'' (see Fig. 4c). Finally, the map for the entire literature collection had ''hippocampal'', ''systems'', ''symbol'', ''rats'', and ''animals'' as the primary concepts in order of frequency (see Fig. 4d).

4.2.2. Themes

''Rats'' was the major theme for the physical navigation map, containing three of the primary concepts and intersecting with the smaller themes of ''cortex'' and ''task'' (see Fig. 4a). ''Navigation'' was contained in its own theme and intersected with the themes of ''cognitive'' and ''cells''. ''Strategy'' was contained in a separate theme and did not intersect with any other themes. For the map on the brain, ''hippocampal'' was the largest theme containing all of the primary concepts, and intersected with the ''place'' and ''spatial'' themes, which both intersected with the ''map'' theme (see Fig. 4b). The ''structure'' theme was positioned in the centre of the

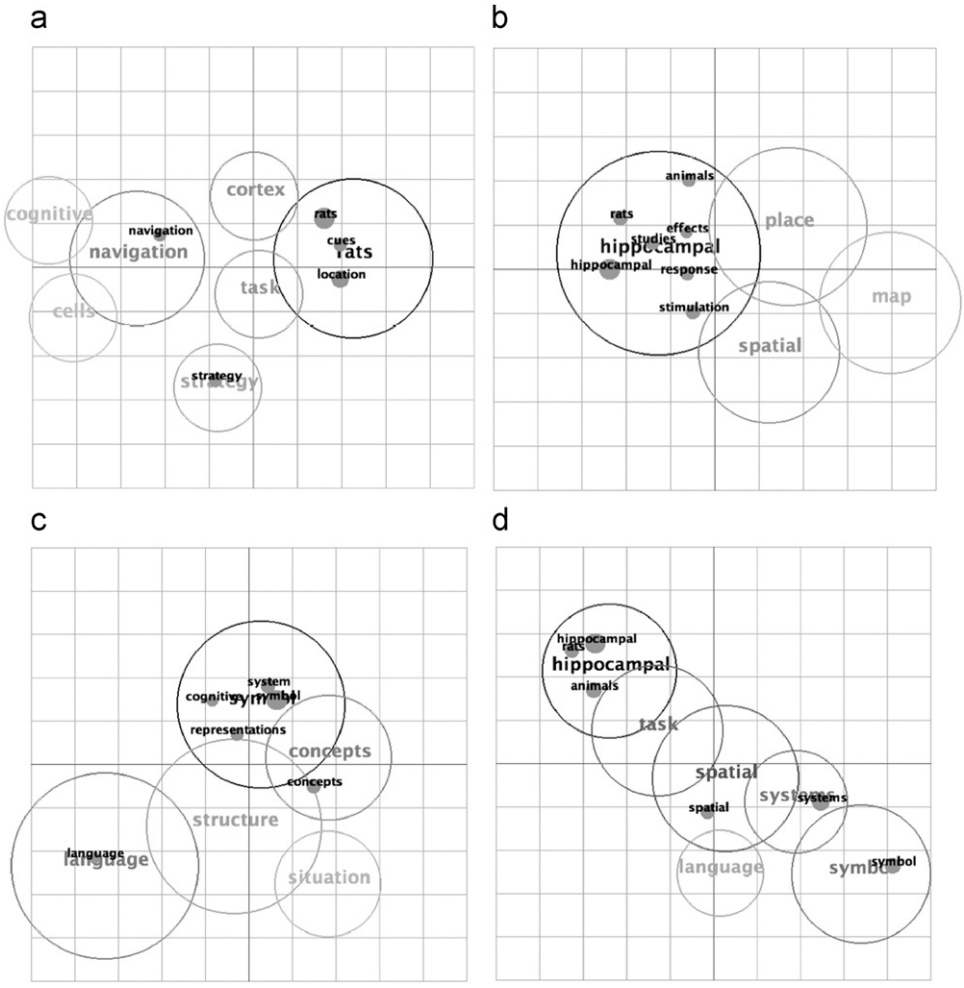

Fig. 4. Leximancer maps for case study 2: (a) map for the subset of the conceptual and spatial navigation literature on navigation with theme size of 37%; (b) map for the subset of the literature on the brain with theme size of 50%; (c) map for the subset of literature on conceptual and spatial mapping with theme size of 46%; and (d) map for all the literature with theme size of 38%.

map for spatial and conceptual mapping and intersected with all the other themes (see Fig. 4c). The theme for "symbol" contained four of the primary concepts, while "concepts" and "language" were contained in their own themes. Finally, the complete literature map had an almost linear structure with "hippocampal" near the top, containing three of the primary concepts (see Fig. 4d). "Spatial" was at the centre of the map, intersecting with "task", "language", and "systems." The primary concepts of "symbol", "systems", and "spatial" were each contained in their own separate themes.

4.2.3. Co-occurrence

All of the primary concepts co-occurred in the physical navigation literature; "rats" co-occurred frequently with "location", while "navigation" co-occurred strongly with "strategy" and "rats." The literature for the brain showed strong co-occurrence for "hippocampal" with all of the other primary concepts, each of which co-occurred less frequently with each other. In the conceptual and spatial literature, "symbol" co-occurred frequently with "system" and "concepts", "representations" with "symbol", "cognitive" with "symbol" and "system", and "system" only with "symbol". All of the primary concepts had some co-occurrence with each other. For the complete literature set, "spatial" showed strong co-occurrence with all of the other primary concepts, while "hippocampal", "rats", and "animals" frequently co-occurred with one another, and "symbol" and "systems" co-occurred frequently.

4.3. Discussion

The second case study investigated the coverage of different sub-domains related by a central topic on a concept map, and what emergent patterns in the concept map can reveal about the selected literature set. The primary concepts from brain literature map overlay the top region of the entire literature map, while the conceptual and spatial concepts overlay the bottom. However, the map for physical navigation also overlaid the top, brain area of the entire literature map—the primary expectation was that each of the sub-domains would occupy different

regions. It was also expected that the central theme for the combined map would be "navigation" as it was considered the unifying concept for each of the three sub-domains; instead "spatial" was the central theme.

To explain the reason for the unexpected positioning of the physical navigation concepts on the entire literature map, the selected literature was examined using Leximancer's text browsing facility. The majority of the selected documents for physical navigation were based on experiments using rats and studying the hippocampus, overlapping with the key concepts in the literature for the brain map. This overlap in the two sub-domains served to reinforce the concept of "hippocampal", increasing both its frequency and co-occurrence in the entire literature map, making it the most significant concept.

"Navigation" was not identified as a primary concept in either the brain literature or in the conceptual and spatial mapping literature; instead it was added to the thesaurus entry for "spatial" in the brain map and into "people" in the conceptual and spatial map. A concept will be added to a thesaurus if it only occurs in the context of a more common concept.

It is likely that "spatial" was central on the combined literature map because of the focus in the literature on spatial tasks in navigation and the spatial aspects of concept mapping. "Spatial" was contained in all three sets of the literature, but was more significant in the conceptual mapping sub-domain (which also included the concept "navigation") and so was positioned centrally between the two regions of the map.

These results show how sub-domains can overlay different regions of a concept map, and that the position may not match preconceptions. Analysing a set of concept maps created using a tool such as Leximancer can highlight the reasons for the discrepancy.

5. Case study 3: a cross-disciplinary topic: the sponge genome project

In the first two case studies, Leximancer was used to examine sets of documents in which the content was well understood by the authors. The third study was motivated by a different kind of information processing challenge. Consider the following scenario: under tight time pressure, prepare for a meeting knowing only the name of the person who called the meeting, and the topic:

> As a bioinformaticist, you are invited to a meeting with the Director of the Sponge Genome project, Professor Bernard Degnan. You don't have long to prepare. All you're really sure about is that you know almost nothing about sponges!

The challenge is not to learn about sponges per se (the topic is actually vast), but to discover the background material that is likely to be relevant to the meeting. In interdisciplinary research, one of the major barriers to communication is the development of a shared lexicon—an understanding of the common terms in each field and how they are used. The challenge in the scenario above is for a technical collaborator to learn a biological vocabulary, prior to an initial introduction to the scope of a project. The goal for this case study was to determine if an automatically generated concept map could guide rapid learning in an unfamiliar domain.

5.1. Finding relevant terms

The case study methodology involved three steps: 1. Retrieve B. Degnan's recent papers using a tool such as Google Scholar (scholar.google.com). 2. Use a content analysis tool to identify the major themes in the selected papers. 3. Look up any unfamiliar terms online.

Google Scholar returned 92 papers by B. Degnan spanning from 1988 to 2006 at the date of the search (16 March 2007). Three papers were selected on the basis of their titles, two recent (2005, 2006), and one highly cited early paper (1995), and the pdfs were downloaded.

Leximancer was used to map the selected papers using standard settings (automatic processing for all stages) with no additional processing (see Fig. 5). The major themes revealed in the study had a central focus on genes, transcription, metazoan and sequence. There were 35 concepts extracted from the text, using default settings. The concept term frequencies showed "genes" was by far the most prevalent, followed by "gene", "genome", and "sequence", ordered by relative counts. As mentioned in the earlier case studies, words and their plurals, such as gene and genes could have been merged, but due to time constraints were left as separate concepts in this case study.

Unfamiliar terms were explored further using the text browsing tool within Leximancer. For example, the term "metazoan" was clearly important (ranked fourth in terms of concept frequency) and had a high co-occurrence with "basal" (19 instances). Usage in the text showed that "basal metazoan" was clearly an important phrase, used in examples such as "a complete genome sequence from the basal metazoan phyla is an obvious goal." Finally, definitions for terms from the ranked concept list were checked online with Wikipedia ("basal" in this context meaning the most ancient in evolutionary terms; "metazoan" which can be roughly translated as an animal; and a "basal metazoan" is a primitive animal that has muscles, a digestive tract, and a nervous system, according to the Wikipedia entry dated 16 March 2007).

The whole process took thirty minutes, including the time to find and download articles using Google Scholar, create and explore the concept map, and find and read definitions for the most common terms. Although a more detailed analysis could have been performed, the main use of Leximancer was to identify the key terms in the set of papers and how they were used together within the available time. The map structure generated by Leximancer showed these relationships clearly and was able to provide sufficient information to gain a basic grounding in the

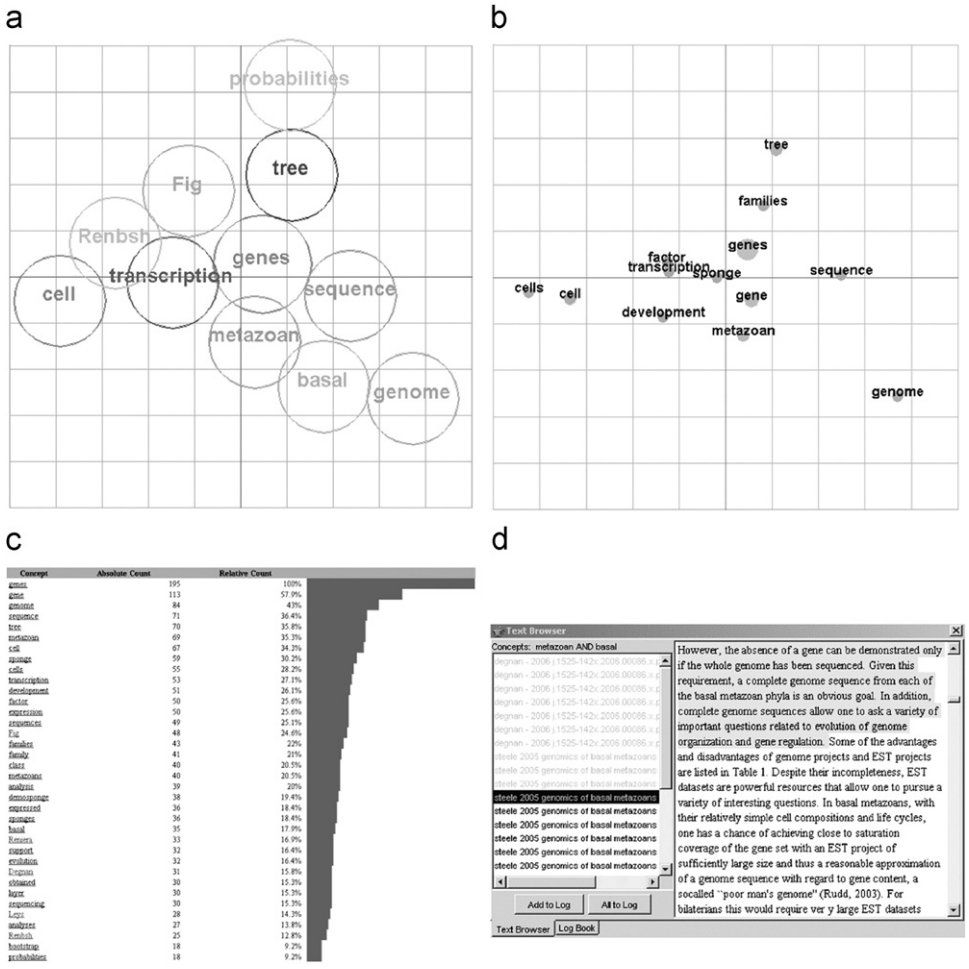

Fig. 5. Leximancer map and concept terms for case study 3: (a) major themes in the sponge genome map at theme size 33%; (b) most common terms at 39% number of points; (c) concept term frequencies; and (d) drilling into the text using text browsing.

domain. Important terms identified using the maps were researched using the web. The suitable application of domain concepts was located using groups of concepts as a phrase, such as "basal metazoan" where "basal" and "metazoan" were both high-frequency terms that frequently co-occurred with one another. In the subsequent meeting, the phrase "basal metazoan" formed the basis of much of the discussion.

5.2. Discussion

The most striking benefit of using a content analysis tool in this case study was the speed of the method when a limited amount of time was available. Without reading the underlying literature in detail, key concepts in the document collection were identified. Sections of the literature were perused for context.

In the first two case studies, time was taken to prune unwanted concepts from the concept list to make a more specific map for the desired analysis. For the third case study, this time was not available, yet the generic map was adequate for performing the desired analysis. To refine the map, concepts such as some word stems could have been merged, and terms related to the papers rather than the content (e.g. "Fig" and non-technical terms) could have been removed. Because of the diverse nature of corpora and different kinds of applications that concept maps may be used for, the analyses and refinement of the map is frequently specific to the needs of the user at the time of the analysis. For the purpose of the case study—rapid learning in an unfamiliar domain—it is doubtful that additional information would be revealed. The utility of the generic map was likely due to the limited set of documents on a single theme.

Various aspects of the information uncovered in the document collection were drawn from different sources: themes, frequency count of terms, co-occurrences of terms and text browsing. Manipulating the theme layout shows the ways in which clusters of concepts can be interrelated. Frequency counts provided evidence of the presence of known terms and directed further analysis for unknown terms where required; however, frequency was not the only criterion for selecting terms for further investigation. Location on the map was also taken into consideration, being a second-order relationship based on the co-occurrence relationships between concepts. Highlighting

co-occurrences of terms is a key to identifying phrases specific to the domain, and can indicate potentially specialist usage of more familiar terms. Examining all co-occurrences of two terms within the text is a fast method of determining how terms are used in context.

Although the concept map and text mining facility were central to the analysis, additional sources of information were required both before and afterward. The Internet was used beforehand for fast identification of potentially relevant papers, and afterward for information on unfamiliar terms identified in the text. These results show that the use of a content analysis tool can aid in rapidly coming to an understanding of an unfamiliar domain and directing attention to critical concepts.

6. General discussion

The goal of this study was to investigate the use of automatic concept mapping for different bodies of literature, showing the focus of a set of related document collections, where they overlap and how they combine. The three case studies revealed different insights into the use of concept mapping and text mining. Real-world scenarios are described where concept maps are used. Although the cases presented are specific, the kinds of analyses and uses for concept maps in terms of these case studies can be extended to the general case.

The first case study created concept maps from a thesis and its background literature. The two maps were used to highlight similarities and differences between the foci of the two different sets of documents. Given the already high level of familiarity with the areas being mapped, it was somewhat surprising to discover the level of additional detail that could be gained from the concept map of the background literature. The comparison between the thesis and the literature maps also prompted a deeper reflection into the major issues in the research field, and showed the extent to which the thesis had effectively covered them.

The second case study was used to explore an unfamiliar set of documents, first by creating separate maps for three separate areas, and then creating one comprehensive map which was used to look for emergent patterns. The combined map showed close relationships between two of the areas, which overlay the same region of space, separated from the third area located in its own region. This structure was unexpected from an initial perusal of the documents, and understanding of the emergent structure was developed through investigation of the different terms in the text itself and in the thesaurus (in this case "navigation" was contained within the theme "spatial"). This study showed that it was not just the software's capability in automatically developing a concept map, but also the ease with which the layout of the map could be used to develop a global understanding of structure, and then the text search facilities could be used to locate specific text segments. The ability to quickly navigate between the local and global structure in the map and documents was used repeatedly in the case studies.

The third case study demonstrated a method for selecting and rapidly acquiring a basic vocabulary in an unfamiliar domain, and its use for cross-disciplinary collaboration. In this study, concept mapping was used in conjunction with other online resources (Google Scholar and Wikipedia) to rapidly identify and find definitions for technical terms most likely to be used by a given researcher. The surprising insight in this case study was how fast one could use automatic concept identification to rapidly identify a set of the most important terms in a given field.

The studies together provided a number of insights into how concept mapping can be used for exploring familiar, semi-familiar and unfamiliar document sets. The Leximancer software uses algorithms that are non-linear and open form for both concept mapping and layout of the resulting map. The algorithm for creating the concept map takes into account the relative frequencies of a given concept when calculating the effect of one concept on another: where one concept mostly co-occurs with a second concept, there will be a strong directed link from the first to the second concept. If the second concept has a different co-occurrence pattern and co-occurs with many other concepts, it will have a relatively weaker link back to the first concept. This approach results in asymmetric networks, and is a major point of difference to other statistical analysis techniques such as HAL (Burgess and Lund, 1997) and LSA (Landauer et al., 1998), which create symmetric networks of concepts. Leximancer also uses non-linear algorithms in the way it displays its concept maps. Non-linear algorithms have both strengths and weaknesses: they can be unstable and give different results as minor changes in information are made to a document set. However, their major strength and the reason why they were chosen for Leximancer is that they can be tailored to provide effective visualisations for human readable maps.

In each case study, active exploration was critical for effective use of the concept maps. In addition to the concept map itself, concepts are collected into groups and given a theme label. The themes are shown as circles covering (possibly overlapping) regions of the maps. These labels and regional groupings add a layer of abstraction to the concept maps and provide additional explanatory power. Leximancer also provides a feature to jump directly to text segments from concept co-occurrences. These hyperlinks enable the user to move rapidly between global scope provided by the concept map, and the details of how the terms are used in context by accessing specific text segments.

In conclusion, automated concept mapping and the associated text mining tools have proven effective for active exploration to identify trends within documents and document collections, for performing differential analysis on documents, and as an aid for learning a new domain. The capabilities demonstrated in the case studies were not due to concept mapping alone, but rather took advantage

of the speed and effectiveness of a software tool that provides a complete solution to text analysis. However, like any tool, it can be used with more or less sophistication—it is clear that active exploration with a concept mapping and text mining tool is a skill. For the novice user, it seems to be most useful in inverse proportion to the user's knowledge of a domain. As users develop in expertise, it can be used to explore subtle aspects of more familiar domains. As with any automated system, skill in its use lies in applying it to problems for which it is suited. The case studies in this paper have demonstrated some such tasks, a key characteristic being the desire to understand both global scope and local details in sets of documents.

Acknowledgements

This research is funded in part by a grant from the Australian Research Council to Janet Wiles and Andrew Smith, and an Australian Postgraduate Award to Paul Stockwell. We would like to thank Daniel Angus, David Prasser, Ruth Schulz, Peter Stratton, and Mark Wakabayashi for comments on earlier drafts.

References

Barrière, C., 2004. Knowledge-rich contexts discovery. In: Tawfik, A.Y., Goodwin, S.D. (Eds.), Lecture Notes in Computer Science, vol. 3060. Springer, Heidelberg, pp. 187–201.

Barrière, C., Agbago, A., 2006. TerminoWeb: a software environment for term study in rich contexts. In: Zheng, W. (Ed.), Proceedings of the International Conference on Terminology, Standardisation and Technology Transfer. NRC, Beijing, China, pp. 103–113.

Benzecri, J.P., 1992. Correspondence Analysis Handbook. CRC Press, New York.

Berelson, B., 1951. Content Analysis in Communication Research. Free Press, New York.

Burgess, C., Lund, K., 1997. Modelling parsing constraints with high-dimensional context space. Language and Cognitive Processes 12 (2/3), 177–210.

Buter, R.K., Noyons, E.C.M., 2002. Using bibliometric maps to visualise term distribution in scientific papers. In: Proceedings of the Sixth International Conference on Information Visualisation (IV'02), London, England, pp. 697–705.

Chalmers, M., Chitson, P., 1992. Bead: explorations in information visualisation. In: Belkin, N.J., Ingwersen, P., Pejtersen, A.M. (Eds.), Proceedings of the 15th Annual ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, pp. 330–337.

Chen, C., 2002. Visualization of knowledge structures. In: Chang, S.K. (Ed.), Handbook of Software Engineering and Knowledge Engineering, vol. 2. World Scientific, Singapore, pp. 201–238.

Dumais, S.T., Platt, J., Heckerman, D., Sahami, M., 1998. Inductive learning algorithms and representations for text categorization. In: Gardarin, G., French, J.C., Pissinou, N., Makki, K., Bougamin, L. (Eds.), CIDKM '98: Proceedings of the Seventh International Conference on Information and Knowledge Management. ACM Press, New York, pp. 148–155.

Hagiwara, M., 1995. Self-organizing concept maps. In: IEEE International Conference on Intelligent Systems for the 21st Century, vol. 1, Vancouver, BC, Canada, pp. 447–451.

Landauer, T., Foltz, P., Laham, D., 1998. Introduction to latent semantic analysis. Discourse Processes 25, 259–284.

Milford, M.J., Wyeth, G., Prasser, D., 2004. RatSLAM: a hippocampal model for simultaneous localization and mapping. In: Tarn, T.-J., Fukuda, T. (Eds.), Proceedings of the International Conference on Robotics and Automation, New Orleans, United States, pp. 403–408.

Mothe, J., Chrisment, C., Dkaki, T., Dousset, B., Karouach, S., 2006. Combining mining and visualization tools to discover the geographic structure of a domain. Computer, Environment and Urban Systems, Special Issue Geographic Information Retrieval 30 (4), 460–484.

Naisbitt, J., Aburdene, P., 1990. Megatrends 2000. Avon, New York.

Novak, J.D., 1990. Concept maps and Vee diagrams: two metacognitive tools to facilitate meaningful learning. Instructional Science 19 (1), 29–52.

Novak, J.D., Gowan, D.B., 1984. Learning How to Learn. Cambridge University Press, Cambridge.

Salton, G., 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading, MA.

Schulz, R., Prasser, D., Stockwell, P., Wyeth, G., Wiles, J., 2008. The formation, generative power and evolution of toponyms: grounding a spatial vocabulary in a cognitive map. In: Smith, A.D.M., Smith, K., Ferrer i Cancho, R. (Eds.), The Evolution of Language: Proceedings of the Seventh International Conference (EVOLANG7). World Scientific Press, Singapore, pp. 267–274.

Schulz, R., Stockwell, P., Wakabayashi, M., Wiles, J., 2006. Generalization in languages evolved for mobile robots. In: Rocha, L.M., Yaeger, L.S., Bedau, M.A., Floreano, D., Goldstone, R.L., Vespignani, A. (Eds.), ALIFE X: Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems. MIT Press, Bloomington, Indiana, pp. 486–492.

Smith, A.E., 2000. Machine mapping of document collections: the Leximancer system. In: Vercoustre, A.-M., Hawking, D. (Eds.), Proceedings of the Fifth Australasian Document Computing Symposium. DSTC, Sunshine Coast, Australia, pp. 39–43.

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

Stockwell, P., 2005. Language games and generalisation grounded in autonomous agents. Unpublished Honours Thesis, The University of Queensland.

Weber, R.P., 1990. Basic Content Analysis, second ed. Sage, Newbury Park, CA.

Yarowsky, D., 1995. Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Cambridge, MA, pp. 189–196.