2010

Can Semantic Mapping Be Used to Model Information Seeking Behaviour in Resource Discovery Systems?

Sophie Cockcroft
University of Queensland, sophie@business.uq.edu.au

Hanna Stelmaszewska
Middlesex University - U.K., hanna_stelm@yahoo.co.uk

Can Semantic mapping be used to model information seeking behaviour in resource discovery systems?

Sophie Cockcroft
UQ Business School
University of Queensland
BRISBANE
Australia
Email: sophie@business.uq.edu.au

Hanna Stelmaszewska
Interaction Design Centre
Middlesex University
UK
Email: hanna_stelm@yahoo.co.uk

Abstract

Locating accurate information on academic topics is a challenge for today's information seekers. Whilst specialist portals exist, users have a marked preference for finding their own information yet are rarely trained in formal information searching using library resources. Further such electronic resource discovery systems are often complex or highly complicated to use. Work is underway to produce a tool to support academic users in accessing relevant information via library sources. Any tool that seeks to assist students and researchers in finding the information they need must be developed in cognisance of their existing workflow. Thus we sought to identify the processes novice and expert users carry out to find information. Using observations with a think aloud protocol, and follow-up interviews we gathered qualitative data on usage and responses to information searching

Keywords

Requirements elicitation, Resource discovery systems, information seeking behaviour, academic libraries, searching

INTRODUCTION

We describe a pilot study whereby we analyse the transcripts of a number of different observation with think-aloud user sessions and interviews. The work presented here is designed to complement more traditional forms of data analysis that have been carried out in this project (Wong et al, 2009) in particular work building on the concepts of Rasmussen and Jensen (1973) who successfully visualised knowledge workers mental procedures in carrying out tasks via computer. Here we ask the questions:

"Can lexical analysis identify user behaviour in the use of electronic resource discovery systems?"

"Can lexical analysis identify the procedural nature of information finding by users of electronic resource discovery systems"

The paper proceeds as follows; first we describe the background to the work in two parts; the use of semantic mapping software referring to a number of prior studies, and issues relating to the use of electronic resource discovery systems. Second this work is positioned in the context of other work on information seeking in library research and education. The methodology is then presented. Finally the results are presented with a discussion and suggestions for further work.

BACKGROUND

Rapid analysis of transcripts containing large numbers of comments and actions may be possible if we are able to semantically map the text contained in the interview or think aloud transcripts. Leximancer is a concept mapping program (Smith, 2005) which can be used to to identify prominent themes in any body of text. It is able to

analyse large bodies of text by tagging, mapping, and mining conceptual data Smith (2005). Leximancer defines collections of words related to a central theme. Families of words relating to a concept are used to create a thesaurus. Some words are more representative of a concept than others, thus Leximancer weights them to ensure evidence of their relationship to a concept is derived from their presence. Leximancer also searches for words that do not occur frequently with the concept to help stabilize the development of the thesaurus and the concept map (Smith, 2005). Leximancer is also able to determine the co-occurrence of concepts in comparable contexts to illustrate the strength of the relationship between the concepts. This is similar to a correlation analysis. A two dimensional map highlighting concept frequency, inter-relationships between concepts and general themes that encompass related concepts is produced by Leximancer. On the map, concepts are clustered according to contextual similarity. Connectedness of a concept to others in the text is illustrated in terms of the size of the concept point. .

The maps are displayed by Leximancer in a way that makes it easy to explore links to related subtext. Every word in the map represents a concept and it is placed on the map based on its relationship to other concepts. Transcripts of the sessions can be analysed to identify unique relationships. In particular, we use an analytical feature called Knowledge Pathway which, when generated across concepts in the data, reveals the most likely path in conceptual space from the start concept to the end concept.

Previous authors have used semantic mapping software for a variety of purposes (see Table 1)

Some of the advantages of using automated lexical analysis tools such as Leximancer include:

- The ability to derive main concept and their relative importance using an objective scientific algorithm,
- The capacity of such tools to compensate for human biases,
- The capacity of such tools to find patterns not obvious to human readers,
- The ability to identify the centrality of concepts,
- The ability to assist in visually exploring textual information for related themes to create new ideas or theories; and
- The ability to assist in identifying similarities in the context in which the concepts co-occur – contextual similarity.

Table 1 Previous lexical analysis studies using Leximancer

Author	Description
Janssen (2008)	In a learning setting they used technologies for semantic lexical analysis to diagnose learner's conceptual development
Davies (2006)	Use of Leximancer to complement results of Nvivo analysis specifically to classify comments otherwise uncategorised by researchers
Grimbeek et al (2004)	Using Leximancer to identify top students' schemata for academic achievement
Miller and Sanderson (2005)	Identification of speech units in video cued recall of patient handover in ICU
Indulska and Recker (2008)	Use of Leximancer as a complementary method for thematic analysis of papers in Design Science to distinguish key factors in papers focussing on methodology and discussion, vs application
Watson et al (2005)	Use of Leximancer to identify unique relationships in pathology case notes
Stockwell et al (2009)	Used Leximancer as a means of rapidly finding key studies in a body of literature

Our work has been informed by all of the above studies in terms of methodology. In particular, as in the work of Indulska and Recker (2008) and Stockwell et al (2009) our study is a complementary one which assesses the value of using semantic mapping to identify themes that are already recognised by other methods. On the other hand it has some elements in common with the work of Janssen et al (2008) which was essentially an educational study, and is described in more detail in the next section.

Models of conceptual development

The work carried out by Janssen et al (2008), set out to analyse learners' conceptual development. It explored differences between novice and expert learners, and positioned them in a domain in terms of their domain knowledge and skill in accessing information. Janssen et al (2008) used Leximancer to compare novice and expert learners them via "think alouds" and tutor notes. They asserted that concept maps capture learners' representations of subject matter structure. They also presented evidence that concept maps are well suited for eliciting knowledge (Nesbit and Adescope, 2006). We see an analogy between their work and the work reported here in that our subjects are also accessing information. Further we also seek to determine the factors that make up expertise in the domain of resource discovery via a variety of systems including:

- library-subscribed such as electronic databases, e-journals portals, (e.g. EBSCO EJS, Emerald, ProQuest), federated search engines, catalogues, e-books and various electronic newspapers (see Adams and Blandford (2002));
- Those freely available on the Internet, including resources such as Google, Google Scholar, Wikipedia, YouTube in digital libraries.

As part of a larger project our data is also being analysed to develop more formal models of this behaviour (Wong et al, 2009). By using semantic analysis the focus is, by definition on the content of what the participants say rather than the procedural nature of the tasks they undertake. This is because semantic analysis has its roots in discovering the meaning and relationship between words. There is, however, an intimate relationship between the two in the form of how the participants conceptualise the search space. Indeed semantic analysis has been used to explore this relationship in the field of education where experts were found to frame learning in a particular medical domain differently than novices (Janssen, 2008), which is summarised in Table 2

Table 2 Adapted from (Janssen, 2008) comparison of Novice and Expert Learners

Expertise Level	Knowledge structure	Learning	Reasoning process
Novice	Networks (incomplete, and loosely linked)	Knowledge accretion, integration and validation	Step by step process
Intermediate	Networks (tightly linked and integrated)	Encapsulation	Big steps (but still one at the time)
Expert	Illness scripts	Illness script for formation	Groups of steps activated as a whole
Experienced Expert	Memory traces of previous cases	Instantiated scripts	Automatic reminding

METHODOLOGY

In our study we implement a pilot scenario for information seeking. Aside from the research questions outlined in the introduction, one of the objectives of this pilot is, following Janssen (2008), to show if automatically generated concept maps can be used to identify patterns of information seeking. We use Leximancer to create concept maps based on input text, which comprises the transcripts of observations with think aloud protocols.

The steps used are as follows:

1. Using as an elicitation method a think aloud protocol, participants are asked to describe the steps they take in addressing the task.
2. This is transcribed into a set of summaries
3. The text of these participant summaries is analysed and compared with the help of Leximancer with an initial set of indicators, e.g. co-occurrence of concepts, relations between them, missing concepts, etc.

With respect to step 1, think aloud protocols involve participants thinking aloud as they are performing a set of specified tasks. Users are asked to say whatever they are looking at, thinking, doing, and feeling, as they go about their task. This enables observers to see first-hand the process of task completion. Observations of the activities undertaken are also recorded.

With respect to step 3, in the pre-processing phase all the defaults are accepted. In the concept identification phase the total number of concepts is set to 100 this is because the data set is conceptually diverse. In the concept seed editor, all identification tags such as MP3 or LP7 are excluded, names of resources such as EBSCO Google and YouTube are selected as seed concepts. The word "Article" is added as a user defined concept as this appears as an endpoint in most tasks. Mis-spelled words such as Wikepiedia are merged with the correct spelling.

There were three user groups; Undergraduate, Postgraduate and Expert. Participants were presented with three tasks which they did in order. Thus there were nine potential units of analysis. For the purposes of this work we focus on a sample of three as described in the results section. The full breakdown is illustrated (participants|(total)) in Table 3. The letters MP, LP and CP reflect the universities from which the participants came. Task descriptions are given in Table 4. It should be noted that the tasks described in Table 4 were defined based on an increasing level of ambiguity making Task 3 more challenging than Task1, and requiring a higher level of ability to find the necessary information.

Nine maps were produced, but we focus on the differences between how Undergraduates, Postgraduates, and Experts approached ***Task 2***. This task was selected as it is a medium level of difficulty.

Table 3 Participants and treatments

Task	Undergraduate	Postgraduate	Expert
1	LP5,LP1, MP8, MP10 (4)	MP2, CP9,CP3,CP7 (4)	MP7,MP11,MP12 (3)
2	LP5, LP1, MP8 MP9, MP10(5)	CP9,CP3,CP7,CP5 (4)	LP7,MP7,MP11,MP12 (4)
3	LP5, LP1, MP8 (3)	CP9,CP3,CP7 (3)	MP7,MP11,MP12, LP7(4)

Table 4 The three tasks presented to participants

Task	Description
1	Product placement is defined as: 'the placement of goods or services in movies and television programs designed to increase brand awareness and brand usage. Find a range of examples from film and television programs, which illustrate product placement 'in action'.
2	The appearance of destinations or locations in films is a form of product placement. There is evidence to suggest from tourism organizations across the world, that when audience see locations in films they are inspired to visit them, so they can 'gaze upon the places where their heroes have been'. Find evidence of film tourism from a range of different film industries to illustrate the impact this may have had on tourism.
3	Imagine that you are the brand manager for a new range of mobile phones for Nokia; you required to produce evidence to demonstrate how you might use the film/television medium as a way of reaching your target audience.

RESULTS

Themes are represented by circles and they form around highly connected concepts. Concepts appear as words. Those in green are proper names. Concepts are conceptually clustered on the map. The brighter the concept the more frequently it occurs in the text. Concepts are linked by a grey network known as a pathway. Using the default setting in Leximancer can result in useful maps, however it is generally more instructive to exploit the semi-automatic elements of the analysis by using various adjustment options to add value. The map output provides the user with a number of sliders to explore the results. Stockwell et al (2009) suggest that initial exploration of the map could involve hiding all of the theme circles to show only the most significant concepts initially, and then sliding the point slider toward 100% to reveal more concepts on the map. They also described an exploration method whereby all themes are shown initially and then are gradually reduced. 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. For the purposes of this study we can only present a static view which has been refined to suit the requirements of the study. In all cases the names of the participants (LP, MP Etc.) were removed from the list of concept seeds at the start of the process. All maps are shown with the theme sizes slider at 50%, the percentage visible concepts was also placed at 50%. In some cases a cluster of grey dots will be observed on the end of a pathway – here a large number of closely related concepts can be observed. All maps presented here are linear maps as they are useful for highly connected networks and they are less sensitive to individual activity. Name-like concepts are singled out for discussion, since most library (internal) and web (external) resources are represented by a name. The other concepts in the maps loosely represent actions or procedures. Knowledge pathways can be imposed, and shown graphically on the map; however these are described at the end of each of the following sections rather than illustrated on the maps, since they obscure the other visual results.

Undergraduate searching approach

In Figure 1 we can see that the name-like concepts "Wikipedia", "Google" and "Library Catalogue" appear. These are green in the original screenshot – in the print version they are a paler colour. Concept frequency, represented by brightness in the screen output, is evident in the associated ranked concept chart in Figure 2, which shows 16 occurrences of the Google concept across the four undergraduate participant transcripts.

Leximancer provides associated extracts from the transcripts, which allows rapid access to relevant comments or observations . Thus we can see that in relation to the "Google" concept, participants' primary approach was to start with Google and enter a search term. These terms included "brand awareness" and "product placement in films and television" (LP1); "evidence of film tourism" (LP5); "Slum dog Millionaire"(MP8); and "film tourism"(MP9). We can also see from the transcript extracts that Google led them to associated resources such as Google Scholar and Amazon.

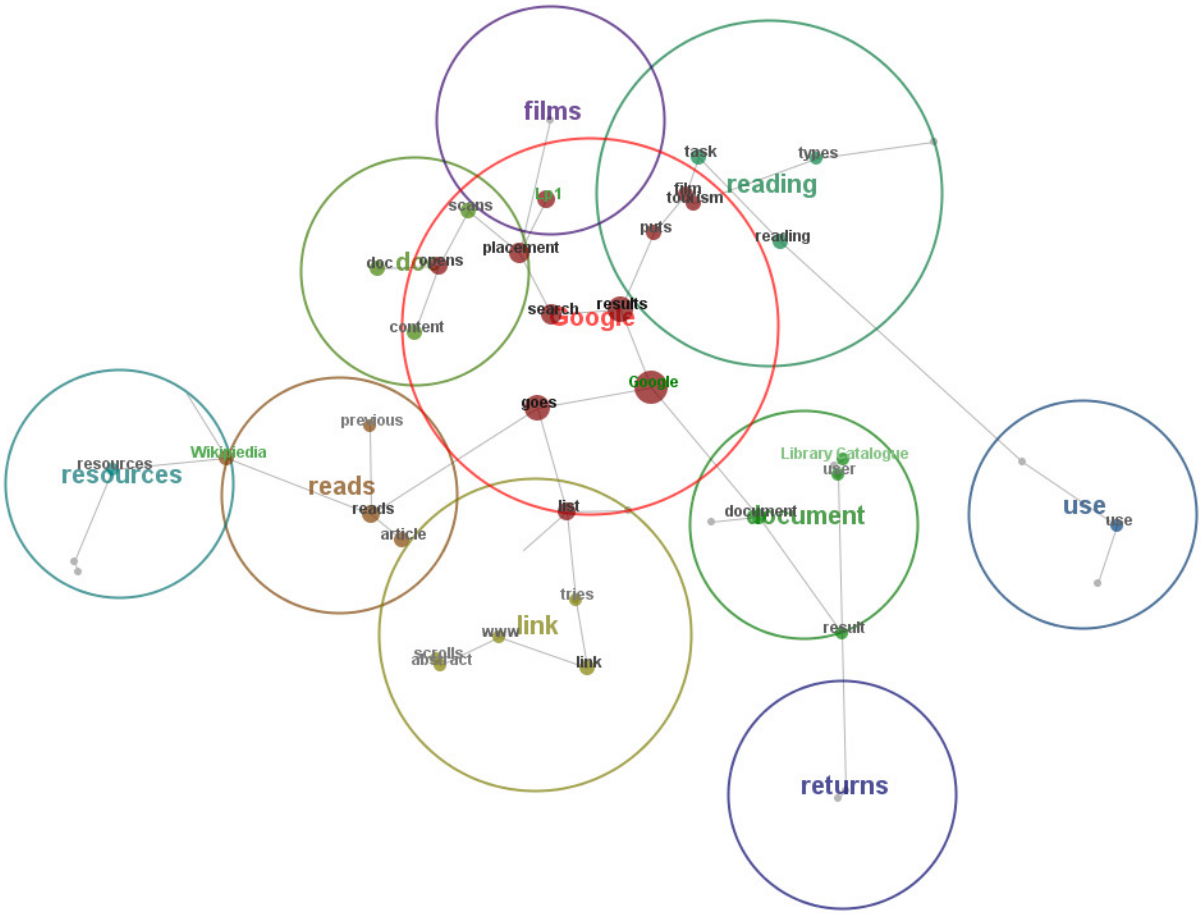

Figure 1 Undergraduate Task 2

Name-Like	Count	Relevance
Google	16	100%
Wikipiedia	3	19%
Library Catalogue	1	06%
Emerald	1	06%

Figure 2 Ranked Name-like Concepts Undergraduate Task 2

Leximancer identified a path in the undergraduate concept map between task and article: **task-puts-Google-placement-list-read-article**

Postgraduate searching approach

In Figure 3, which illustrates the postgraduate searching approach, we also see Google as a prominent searching approach. With Proquest featuring as a secondary approach .(frequencies are given in Figure 4). Participant MP2 commented "Sometimes I also look for the authors if I'm familiar with, and maybe if you type something but you can find something related, and you get a link for another website". In relation to this, and other comments, Leximancer has identified an "authors" theme. "Athens" is a proprietary search portal at one of the universities involved.

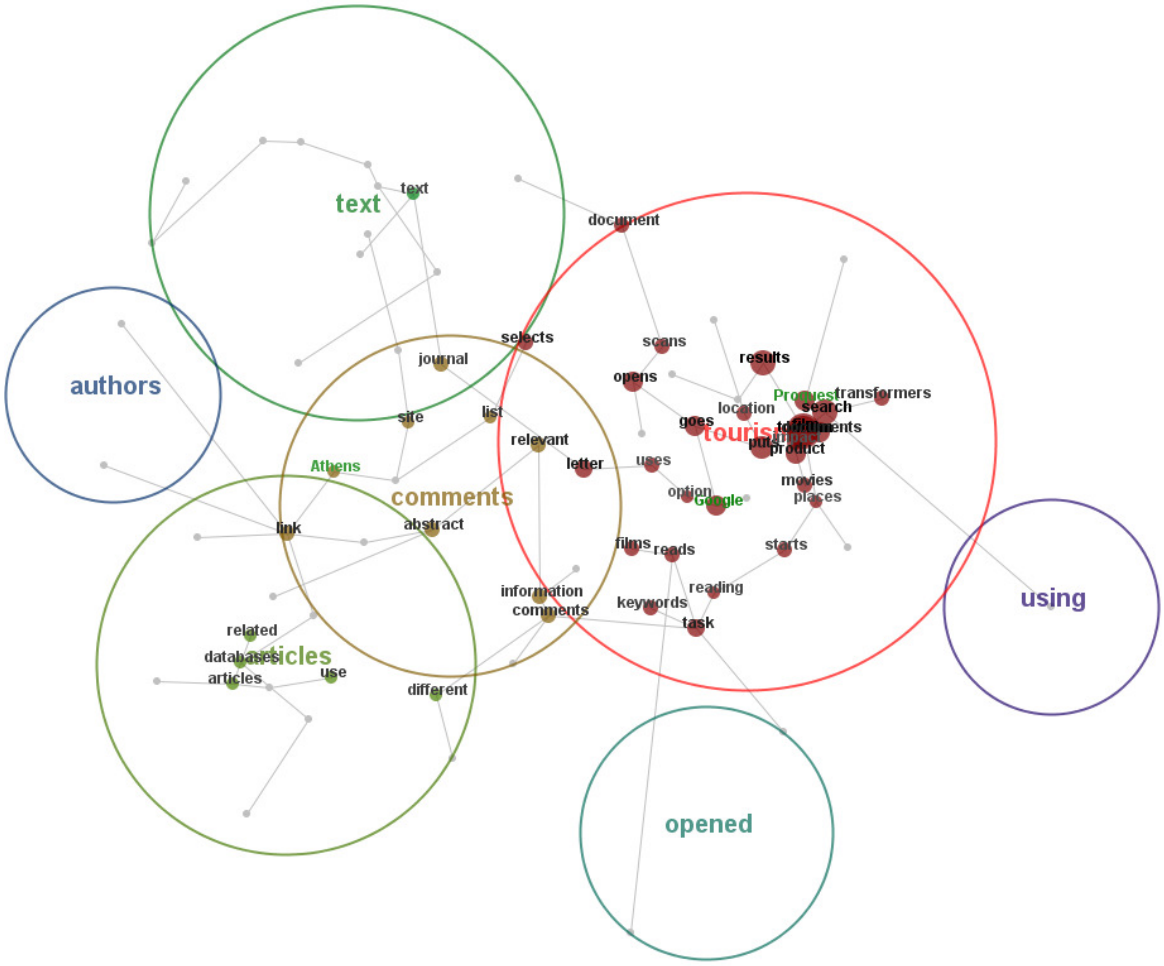

Figure 3 Postgraduate Task 2

Name-Like	Count	Relevance
Google	18	62%
Proquest	10	34%
Athens	6	21%
Multisearch	4	14%
Transformers	2	07%
Journal Of Travel	1	03%

Figure 4 Ranked Name-like Concepts Postgraduate Task 2

Leximancer identified a path in the postgraduate concept map between task and articles: **task-comments-different-use-databases-articles.**

Expert Searching approach
All four expert participants began their search by accessing journals. This was generally via a database such as JSTOR or EBSCO. The types of activities carried out are illustrated in the transcript extract below:

"She goes to ***E-Journals*** and scrolls the list of journals. She then selects a link to **JSTOR** collection" (LP7); "He goes back to the ***Emerald*** and refines the search '***Evidence*** of film industry with local destination' 'all fields' – 34 results."(MP11); " –Oh, this is a **SAGE** journal. SO **EBSCO** doesn't have access to **SAGE**" (MP12); **"**When the page is opened she reads the abstract and decides to view the full text. However, she doesn't use the '***Full text***' option available on the article page but she goes to the ***Journal*** of vacation marketing page from ***E-journals*** in library resources and selects the lick to '***Sage Publications***' When the journal page is displayed she selects '2003' , then 'April' and scrolls the page content of that issue"(MP7).

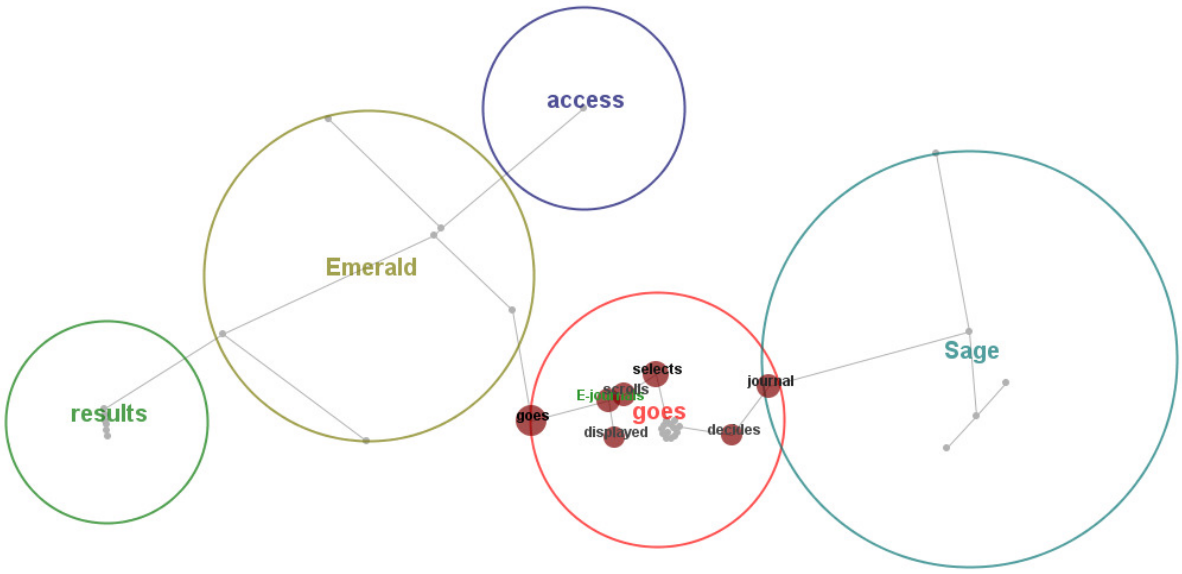

Figure 5 Expert Task 2

Name-Like	Count	Relevance
E-journals	3	60%
Emerald	3	60%
Sage	3	60%
Journal	1	20%

Figure 6 Ranked Name-like Concepts Expert Task 2

There is a large amount of information that cannot be seen on the map in Figure 5 due to overlapping concepts around results and "goes". In particular around results we find the concepts "tourism", "articles" and "search".

Leximancer identified a path in the expert concept map between task and articles: **task- E-journals-scrolls-journals- link- Emerald-film– destination-results tourism- articles**

DISCUSSION

Earlier completed research on this data (Wong et al, 2009) found that the undergraduate group were more inclined to use external resources such as Wikipedia, YouTube and Google. The postgraduate group were more likely to use Google scholar and both postgraduate and experts were more likely to use internal (library) resources. Where they did pick a library resource EBSCO, ProQuest and Emerald were the most common choice. The concept maps produced in this work are a represent a crude analysis in comparison to the detailed qualitative analysis that has been completed as part of the Ubird project. However, high level content information is apparent for example the resources used. The more frequent use of external resources by

undergraduates is supported, but there appears to be a task related preference – with the undergraduates and postgraduates attempting to use internal resources more frequently in task 2. By contrast, and in agreement with the qualitative study, internal resources appear in all tasks carried out by experts.

Our attempts to find a path through the data using knowledge pathway have been more problematical. For this analysis we used "task" as the start term and "article" as the end point. Leximancer provides further analysis in terms of text extracts to see how this path evolves, and the strength of the path as a percentage explained analogous to an R-value. If this method were to be used again the interviewer and transcriber could include consistent words that indicate progression throughout the task to assist in this approach. Clearly the maps alone do not give the full story. Experts brought more experience and knowledge of existing tools to their use of library resources. For example at least one expert user referred to journals in their search which they got to directly (by knowing the name of the journal) this is known in part because Leximancer provides the ability to further explore given concepts and their relationships to other concepts via extracts of text as described above.

Early results suggest tentatively that in answer to the original research questions, some level of user behaviour can be measured using lexical analysis, but it is not an appropriate tool for exploring the procedural aspects of the tasks.

Further work could include expansion of the work to look at post experiment interview data as well, and the use of key words in the observation protocol.

REFERENCES

Adams, A. and Blanford, A. 2002. Digital libraries in academia: challenges and changes. In E-P Lim, S. Foo, C. Khoo, H. Chen, E. Fox, S. Urs and T. Costantino (Eds.) Digital Libraries: People, Knowledge, and Technology pp 392 - 403. Springer LNCS 2555.

Davies,I. Green, P. Rosemann, M. Indulska, M and Gallo, S. 2006. How do practitioners use conceptual modeling in practice? Data & Knowledge Engineering, 58 pp 358-380.

Ellis, D. and Haugan, M. 2004. "Modelling the information seeking patterns of engineers and research scientists in an industrial environment" Journal of Documentation 53(4) pp 384-403

Grimbeek, R., Bartlett, B. and Loke, K.K.2004. "Using Leximancer to Identify Themes and Patterns in the Talk of Three High-Distinction Students" Peter Grimbeek, Brendan Bartlett, & Kit-Ken Loke Griffith University Institute for Educational Research report no. 10072

Hubert Vogten and Rob Koper 2008. "Towards a learning path specification" Int. J. Cont. Engineering Education and Lifelong Learning, Vol. 18, No. 1 pp 77-97

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. pp 1-14

Janssen,J, Janssen,A. Vogten, H. Koper, R. 2008. "Towards a learning path specification"… International Journal of Continuing Engineering Education and Life Long Learning Issue: Volume 18, Number 1 pp:77 - 97

Kuhlthau, C., 1988. "Longitudinal Case Studies of the Information Search Process of Users in Libraries." Library and Information Science Research (July-September) 10(3) pp. 257-304.

Marchionini, G 1995. "Information Seeking in Electronic Environments" Cambridge University Press

Miller, A and Sanderson, P. 2005. "Clinical information use by medical and nursing staff in the ICU: Outcomes of a coded tables analysis" Proceedings Of The Human Factors And Ergonomics Society 49th Annual Meeting pp 984-988

Nesbit, J. C., & Adescope, O. (2006). Learning with concept and knowledge maps: A metaanalysis. Review of Educational Research 76(3), 413–448.

Rasmussen, J. and A. Jensen 1973 A study of mental procedures in electronic troubleshooting. Riso Technical Report No. Riso-M1582. Riso National Laboratory, Roskilde, Denmark, 1973.\\

Smith, A.E., 2005 "Leximancer Manual" Unpublished Manuscript, The University of Queensland, Australia 2005

Stelmaszewska, H., Wong, D., Bhimani, N., and Bam, Balbir 2009"Information seeking in electronic resource discovery systems" Working Paper University of Middlesex.

Stockwell, P, Robert M. Colomb, Andrew E. Smith, Janet Wiles 2009. Use of an automatic content analysis tool: A technique for seeing both local and global scope. Int. J. Hum.-Comput. Stud. 67(5): pp 424-436

Watson, M. Smith, A. Watter, S. 2005 "Leximancer concept mapping of patient case studies" Lecture notes in Computer Science 3683 pp 1232-1238

Wong,W. Stelmaszewska,H., Bhimani , N. Barn S. and Barn, B. 2009 ".User Behaviour in Resource Discovery: Final Report" Retrieved 14 March, 2010, from http://www.ubird.mdx.ac.uk/wp-content/uploads/2009/11/ubird-report-final.pdf

ACKNOWLEDGEMENTS

We acknowledge the support of Middlesex University for hosting one of the authors during the research and writing of this paper. The UBiRD Project, on which this work was built, was funded by JISC, Grant Reference No. CSSERSA2 / SERV ENHANCE. We would like to thank the participants from the London School of Economics, Cranfield University, and Middlesex University, who took part in this study, and the librarians who freely gave of their time to provide us with access to the resources.

COPYRIGHT