Qualitative Data Mining and Knowledge Discovery Using Leximancer Digital Software

Charles Kivunja
University of New England, School of Education, Armidale, New South Wales, Australia
Email: ckivunja@une.edu.au

***Abstract*—It is the nature of qualitative research data to be collected in large amounts of interview narratives or conversations about human behaviour that cannot be analysed using conventional quantitative computer software. Such data need to be mined using qualitative digital software that can make sense of letters and words rather than numbers and equations. One such software that appears to be very efficient at mining qualitative data is called Leximancer and is the topic of this paper.**
Thanks to researchers at the University of New England and the University of Queensland in Australia, Leximancer can be used to mine large amounts of qualitative textual documents, extract information at super-electronic speeds and display the results visually in a graphic organiser of the contents generally called a Concept Map. The researcher is able to mine the data deeply to discover the meaning embedded in its digital structures through conceptual (*thematic*) analysis as well as relational (*semantic*) analysis in a manner which can be a great time saver for the researcher.

***Index Terms*—qualitative data mining, conceptual analysis, relational analysis, knowledge discovery**

I. Introduction

Qualitative researchers usually collect large amounts of data about human behaviour which cannot be analysed using the conventional quantitative statistical packages. In the absence of digital qualitative software, the researcher has to visually read all the information transcribed from such interviews, code it and try to discover the meaning embedded in the data. This is where Leximancer [1] qualitative data miner can be a great time saver and a close companion of every qualitative researcher.

Leximancer is digital software that can be used to mine the content of large volumes of qualitative data from interviews or documents at very fast speeds, extract information and display the results visually in a bird's eye view of the contents. The results of the analysis are displayed in what is called a Concept Map although it shows more than just the concepts in the data.

II. Software Functionality Description

A great advantage of this approach is that Leximancer inductively extracts the concepts for you the researcher and there is no need for the researcher to predetermine which themes or concepts are to be coded for. This can be seen as a clear advantage which decreases the researcher's possible subjectivity in data analysis. Moreover, the themes and concepts are grounded in the data [2]. The Concept Map shows more than just the concepts as it also shows the themes within which the concepts are grouped, the interrelationships among those concepts as well as the relative and absolute frequencies of the concepts. Leximancer mines the words in the data and if a number of words move in contextual space with other words often enough, Leximancer performs conceptual (*thematic*) analysis as well as relational (*semantic*) analysis [3] and produces concepts and themes through a complex process which can be very simply illustrated as shown in Fig. 1.

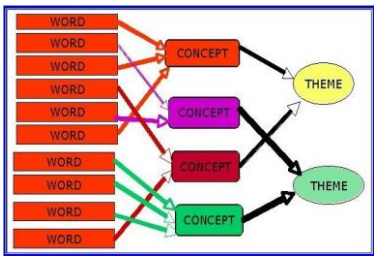

Figure 1. Leximancer's Concept Map development process

The primary graphic produced from this process is called a Concept Map and is intended to simplify a rather complex process that Leximancer goes through involving seven stages illustrated in Fig 2.

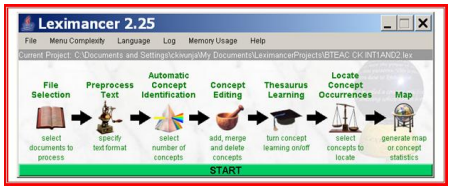

Figure 2. Leximancer's 7 Processing Stages

The first stage is the File Selection stage. This is the stage at which the qualitative data are uploaded into the software. The second stage is the Pre-processing stage. In this stage, Leximancer examines the data to see if it is

Manuscript received Dec 13, 2012; revised Feb 20, 2013.

doi: 10.12720/lnit.1.1.53-55

'qualitative in nature'. By qualitative in nature I mean that the data are not weighted towards numbers. That is to say that they consist of mainly words as used in natural language rather than numerical, counting values. This is the type of research data contained in interview transcripts. During the Pre-processing stage the data are converted into a format that Leximancer recognises and can mine to discover knowledge embedded in the words.

As illustrated in Fig. 2, the third stage is called Automatic Concept Identification. Here Leximancer identifies "seed words" for each concept. By seed words I mean the starting building blocks for the definition of each concept. For example, in the story of Cinderella and the Prince, which is the tutorial supplied with the software, seed words include Cinderella, princess, prince and golden slipper. In my research on students learning in New South Wales multicampus colleges, the seed words included college, multicampus college, students, learning and other such key words as can be seen in the Concept Map in Figure 3. Thus, seed words may be any single or compound words that are the central key words in the textual space of the data. Leximancer mines for these words as they naturally emerge in the data and if a number of words move in textual space, often enough, in association with another word, then Leximancer constructs a concept. The concepts are then grouped into themes as shown in Fig. 3. As shown, in the Theme College in my data referred to above, the concepts included schools, multicampus, colleges, college and contexts.

This stage is said to be similar to a Grounded Theory [4] approach because the concepts are discovered through mining the data for relational meaning. As suggested earlier, this is seen as a significant advantage in qualitative research because the existing data and the relationships among them determine what concepts are important rather than the researcher pre-determining what concepts should be found in the data.

In Fig. 3, the primary circles are called Themes and the blobs inside each circle are called Concepts. This, Concept Map consists of thematic circles and concepts which Leximancer generated from my data which investigated a new multicampus college's impacts on students, their learning and the questions that stakeholders were asking about the new college. The Figure also shows a back ground of a grid. This can be used to investigate the centrality of Themes and Concepts.

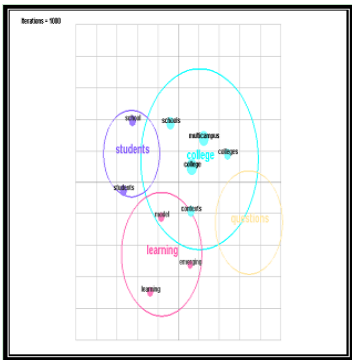

Figure 3. Leximancer Concept Map from author's analysis of research data on Multicampus Colleges in NSW, Australia.

The most prominent feature of the Concept Map are the multi-coloured circles. As said above, these circles represent the key Themes in the data rather than concepts, even though this is called a Concept Map. The circles are shown in different colours for a good analytical reason. The brighter the colour, the higher the dominance of the theme in the data.

Concepts are enclosed within the circle representing the theme to which they belong. They are shown as multi-coloured blobs. Again the colour coding serves the same purpose as that described above for themes. As one would expect, the most prominent concept within a theme, is given the same name as the theme. Other concepts that occur within that theme are then clustered around the key concept within the theme. Detailed information about each concept, within and across themes can be investigated by working a few navigation buttons which enable you to call up relationships among concepts and supporting evidence from within the textual data.

In addition to Themes and Concepts, this stage also generates statistical data which can be used to analyse the Relative and Absolute Frequencies of all the concepts.

The Concept Editing stage 4, gives the researcher the opportunity to merge concepts that are closely interconnected, if they need to, to delete any that the researcher feels crowd the Concept Map unnecessarily or others that may be regarded as "nonsense concepts". An example of the later could be respondents' use of words such as aaah, uuh, well, yaah; repeatedly.

In the Thesaurus Learning stage 5, Leximancer mines the data through thousands of iterations which examine the clusters of words which move together with key terms in the data. As said earlier, if some words are observed to move frequently in association with certain term in the data, and less frequently with others, Leximancer learns that the words that travel with such a term configure the profile of that term and define that term as a key concept in the data.

In this process, Leximancer engages in content analysis and searches for contextual evidence. The evidence is weighted and when the weighted accumulation of evidence reaches a learning threshold (e.g. 10 words moving together in textual space), then Leximancer learns the set of words that she uses to define each concept. These words are then included in the Concept Thesaurus in the software. You can influence Leximancer's learning of the Concept Thesaurus by decreasing or increasing the number of words that should be used to define a concept. Increasing the learning threshold lowers the relevancy threshold. Lowering the former, increases the latter.

Stage 6 is called Locate Concept Co-occurrences. This is the stage during which the circles are dispersed on the map according to the degree of co-occurrence of the themes in the data. Intersecting circles show co-occurrence of themes. Circles that are further apart show themes that don't move together in the data. The grid helps in investigating such co-occurrence. This is followed by the final stage 7 which produces the primary product which, as said earlier is called the Concept Map.

III. CONCLUSION

It is hoped that this Conference will create opportunity for the demonstration of the functionality of Leximancer software described in this paper so that qualitative researchers might increase their capacity for qualitative data mining to enhance their knowledge discovery of the contents of their research data.

REFERENCES

[1] A. Smith. (2008). Leximancer Version 2.25 qualitative Data Software. Retrieved August 28. [Online]. Available: http://www.leximancer.com

[2] Y. Lincoln and E. Guba, *Naturalistic Inquiry*, Thousand Oaks: Sage, 1985.

[3] M. B. Miles and A. M. Huberman, *Qualitative Data Analysis: An Expanded Sourcebook*. 2nd Edition. Thousand Oaks, CA: Sage, 1994.

[4] B. G. Glaser and A. L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research***,** New York: Aldine de Gruyter, 1999.

Dr. Charles Kivunja is a Senior Lecturer in Leadership and Pedagogy in the School of Education, at the University of New England, where he won the Award for Excellence in Teaching in 2009 and the Excellence in Unit Development Award in 2012. He gained his PhD in Leadership and Pedagogy from the University of Western Sydney-Australia. He holds three Masters degrees in Economics and Management one from each of the University of Sydney, University of Western Sydney and the University of Nairobi. His first degree was a Bachelor of Economics with Honours and a Diploma in Education.

He lectured at the Australian Catholic University in Sydney before taking up his current appointment. His current roles include coordination of the doctoral Unit on Leadership and Culture in the Workplace, the teacher preparation Unit on Pedagogy and the Course for Master of Teaching (Primary). He was Leader of an international, education capacity building research partnership, funded by the British Council (DelPHE), involving the University of New England, the University of Zambia and Kyambogo University in Uganda, to boost the use of multigrade pedagogy and strategies in formal and non-formal education in Sub-Saharan Africa. In his current teaching and research, he has embedded cutting-edge technologies into constructionist pedagogy including intensive use of social media tools such as *Google + Discussion Circles*. He is a pioneer in investigating the functionality of Leximancer software, which is a versatile qualitative data miner. He is the Manager of Leximancer software applications at the University of New England. His research and many publications include the use of Leximancer in data analysis.

