Full Paper

Introduction

A version of the Course Experience Questionnaire (CEQ) has been included in the Graduate Careers Council of Australia national survey of university graduates from 1993 onward. For all CEQ items, respondents are asked to express their degree of agreement or disagreement using a five-point response measure. These response data are aggregated to form a series of numerical 'scales' for reporting. In addition to the quantitative response items noted above, the CEQ also includes an invitation to respondents to write open-ended comments on the best aspects (BA) of their university course experience and those aspects most needing improvement (NI). These responses provide a rich source additional qualitative information that can help in understanding what students had in mind when agreeing or disagreeing with the numerical response items (Zaitseva, Milsom, & Stewart, 2013).

The collection of textual data in large-scale surveys is commonplace, due to the rich descriptions of respondent experiences they can provide at relatively low cost. However, historically these data have been underutilised because they are time consuming to analyse manually, and there has been a lack of automated tools to exploit such data efficiently (Bolden & Moscarola, 2000; Jackson & Trochim, 2002). The computer-based analysis and visualisation of textual data goes by various names, including lexical analysis (Bolden & Moscarola, 2000), concept mapping (Jackson & Trochim, 2002; Zaitseva et al., 2013), text mining (Ishii, Suzuki, Fujii, & Fujiyoshi, 2013; Minami & Ohura, 2013; Richardson, 2003), and text analytics (Hu & Liu, 2012; King, 2009). We will use the latter term as the general name for describing, "… a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation." (Hu & Liu, 2012, p. 388) A typical visualisation output from text analytics software is a two-dimensional (2D) chart that identifies key words or themes in the source text, indicates the relative frequency or importance of those words/themes, and represents in 2D some aspect of the relationships between the words/themes. There are many published examples of text analytics applied to open-ended text data, including survey comments, but case studies using student evaluation of teaching data are much less common. Richardson (2003) used the Leximancer software package to analyse 46,000 CEQ comments received at an Australian university to produce a visualisation identifying key themes present in the BA/NI student comment data. Zaitseva et al. (2013) also used the Leximancer software package to analyse several thousand National Student Survey comments (from final year students) received at a UK university, as well as comparable comment data from first and second year students, to identify key themes in the student comments, and to examine how they differed between year levels.

The Faculty of Science, Engineering and Built Environment (SEBE) at Deakin University in Australia is made up of four academic Schools, including Engineering, and receives a relatively large volume of CEQ student comment data annually. Historically these data have been difficult to interpret in a meaningful and timely way without extensive manual coding of the open-ended comment text. Text analytics approaches offer analysis methods that result in visual representations of comment data that highlight key individual themes in these data and the relationships between those themes. This paper reports on a research project to develop and evaluate text analytics methods for the visual analysis of CEQ open-ended comment data from the SEBE Faculty at Deakin University, and to identify the important themes present in these CEQ student comment data. The project aimed to visualise the themes in these CEQ comment data at the following levels: i) whole of Faculty; ii) intra-Faculty/inter-School; and iii) individual School. Via a case study of the analysis of an annual set of CEQ student comment data presented here, we describe in detail the process developed and offer it as a methodology that could be used by others.

Method

As required by institutional ethics processes, exemption from ethics approval was obtained for the use of a de-identified annual set of CEQ comment data for the Deakin University SEBE Faculty. The text analytics software package KH Coder (Higuchi, 2014; Ishii et al., 2013; Minami & Ohura, 2013) was used to analyse an annual CEQ open-ended comment set for the SEBE Faculty. KH Coder was selected as it is free and provides a range of analysis and visualisation options described below. KH Coder supports the use of a dictionary of 'stop words', that is, words to be ignored in any analysis of the text (Hu & Liu, 2012). Common English words and parts of speech, such as 'I', 'a', 'am', 'be', 'my', 'the', etc., add little to the analysis, and their relatively high frequency often masks the words/terms that are actually of significance (Bolden & Moscarola, 2000). A stop word dictionary was developed based on the example English stop word dictionary supplied with KH Coder, after inspection to remove any words likely to be relevant in the context here, such as 'computer'.

A second issue that can mask the significance of words/terms in text analytics is the presence of inflected and/or derived forms of words, for example, a key root word such as 'write' may also be present in the source text as 'writing', 'wrote', 'written', etc. KH Coder implements 'stemming' to consolidate inflected and derived words into their root form. KH Coder supports stemming using Porter's 'snowball' algorithm (Hu & Liu, 2012), or via 'lemmatisation', which first attempts to break the source text into standard parts of speech prior to consolidating words into their root forms (Bolden & Moscarola, 2000). Here we use stemming via lemmatisation based on English parts of speech (nouns, proper nouns, adjectives, verbs, etc.). In text analytics a 'unit of analysis' is required, that is, what is the smallest elemental grouping of text upon which the analysis will be based. KH Coder supports sentences and paragraphs as units of analysis. In our data, each student comment is represented as a paragraph in a text file. It is individual student comments that are of interest here, so we choose the unit of analysis as paragraphs. KH Coder supports a range of text data analysis and visualisation methods – the two that we employ here are multi-dimensional scaling (MDS) and the co-occurrence network (CON).

Generically, MDS computes a measure of 'similarity' (or conversely 'distance') between all pairs of text terms, then seeks a representation (visualisation) of the terms in the least possible number of dimensions, such that original similarity/distance values between all term pairs are shown with the least error possible (Namey, Guest, Thairu, & Johnson, 2007). While MDS can be implemented manually (Jackson & Trochim, 2002), large data sets and many distance and dimensional reduction algorithms are best suited to computer implementation. The error in the resultant visualisation is reduced as more dimensions are used, however using more than two dimensions makes the visualisation hard to display and interpret visually (Namey et al., 2007). KH Coder supports a number of distance measures and dimensional reduction techniques – here we use the Jaccard distance measure (Hu & Liu, 2012) and the Kruskal distance scaling method for dimensional reduction (Chen & Buja, 2009). KH Coder can perform MDS in one, two or three dimensions and visualise the result – here we use 2D MDS as a trade-off between the fidelity of the representation of distances and the ease of interpretation of the visualisation. Words/terms clustered close together in the resultant MDS visualisation are found more frequently close together in the source text, and may reveal key themes in the student comments. Based on specifying the minimum frequency of occurrence of a term for inclusion in the MDS analysis and visualisation (Zaitseva et al., 2013), terms appear as circles/bubbles in the plot, and it is possible to configure the plot to indicate the relative frequency of terms by the relative size of their bubble. It is possible to vary the minimum frequency of occurrence of a term, to examine the impact on the analysis. KH Coder provides the exploratory facility to identify by group colour different numbers of clusters in MDS visualisations based on dimensional similarity.

Co-occurrence refers to the presence of two (or more) words/terms in the same unit of analysis (Namey et al., 2007) – here we are interested if the same word/term pairs/groups frequently co-occur in student comments. KH Coder uses the Jaccard distance as a measure of co-occurrence for term pairs. Based on specifying the minimum frequency of occurrence of a term for inclusion in the CON analysis and visualisation, terms appear as circles/bubbles in a network plot based on the Fruchterman and Reingold (1991) layout algorithm. Frequently co-occurring terms in the visualisation are connected by lines. It is possible to configure the plot to indicate the relative frequency of terms by the relative size of their bubble, and to indicate the relative frequency of co-occurrence of terms by the relative thickness of the line connecting their bubbles. KH Coder provides the exploratory facility to apply a range of colour coding schemas to emphasise different network features. KH Coder provides a key-word-in-context (KWIC) concordance feature that can identify the locations in the source comments of phrases that contain one or more specified keywords within a specified distance of each other (Bolden & Moscarola, 2000). Based on identifying pairs/groups of terms appearing in MDS and CON visualisations that are of interest to investigate further, the KWIC concordance feature allows these term groupings to be located in their original comment context for consideration.

The BA student comment set for the entire Faculty was visualised as a MDS plot, choosing the minimum term frequency to be included in the analysis such that the resultant visualisation contained approximately 50 terms (Bolden & Moscarola, 2000). The same comment set was visualised as a CON plot, with the number of terms to include specified to be the same as the number ultimately included in the MDS plot. The resultant MDS and CON plots were examined to identify key themes emerging, especially themes indicated by both forms of visualisation. The KWIC concordance was used to interrogate the terms related to the identified themes in the original context of the source comment set, to see if there were consistent messages being presented by students. This visualisation/ interrogation process was repeated for the NI student comment set for the entire Faculty. Each individual student comment in the BA comment set for the entire Faculty was tagged/prepended with an identifier indicating the owning School for the program of study to which the comment was related. The visualisation/interrogation process was repeated, resulting in new all-Faculty MDS and CON visualisations that included a locus point bubble for each School, positioned within all of the term bubbles according to the analysis and layout rules for the particular type of visualisation (Bolden & Moscarola, 2000). This intra-Faculty/inter-School form of visualisation provided a view on where the Schools sat in relation to each other and all the of included comment terms, within the resultant 2D space of the particular type of visualisation. This School-based tagging and visualisation/ interrogation process was repeated for the NI student comment set for the entire Faculty. Finally, the visualisation/interrogation process was repeated for the individual BA and NI student comment sets for each of the four Schools in Faculty separately, to yield MDS and CON plots that provided a more detailed/focussed view of comment themes for the unique context of each School.

Results and Discussion

The annual CEQ open-ended comment set for the SEBE Faculty at Deakin University used here contained 482 BA and 458 NI comments, containing 13,571 words, from 513 individual student respondents across 55 separate academic programs. For the period in question, the overall Deakin CEQ response rate was close to the median of all Australian universities.

Whole of Faculty level

Figure 1 presents the MDS visualisation for BA comments for the entire Faculty. For practical readability of the visualisation, a lower limit has to be chosen for words to be included in the analysis – here a limit of word frequency of 13 or greater resulted in 51 terms being included in the analysis. Acknowledging that the identification of clusters in the visualisation is indicative rather than definitive, 12 separately coloured clusters are shown. Figure 2 presents the CON visualisation for BA comments for the entire Faculty based on the same set of 51 terms. One notable feature present in both Figure 1 and Figure 2 is relatively large bubbles for the terms 'practical' and 'work' that are closely associated (MDS) and closely connected (CON). The KWIC concordance feature was set to use 'practical' as the primary search term, in conjunction with 'work' appearing within five words either to the left or right. Table 1 presents the KWIC concordance summary of source comment text entries meeting the search criteria. It can be seen that students regularly reported the value of practical work in their studies. Other notable term pairings apparent in Figure 1 and Figure 2 include: 'good' and 'lecturer'; 'learning' and 'environment'; 'interesting' and 'subject'; and, 'research' and 'project. The whole of Faculty visualisation process was repeated for the NI comment set. Together, this set of four visualisations provides an overview of the key themes/issues reported by students responding to the open-ended comments section of the CEQ that are most relevant for the Faculty-level teaching and learning administrators.

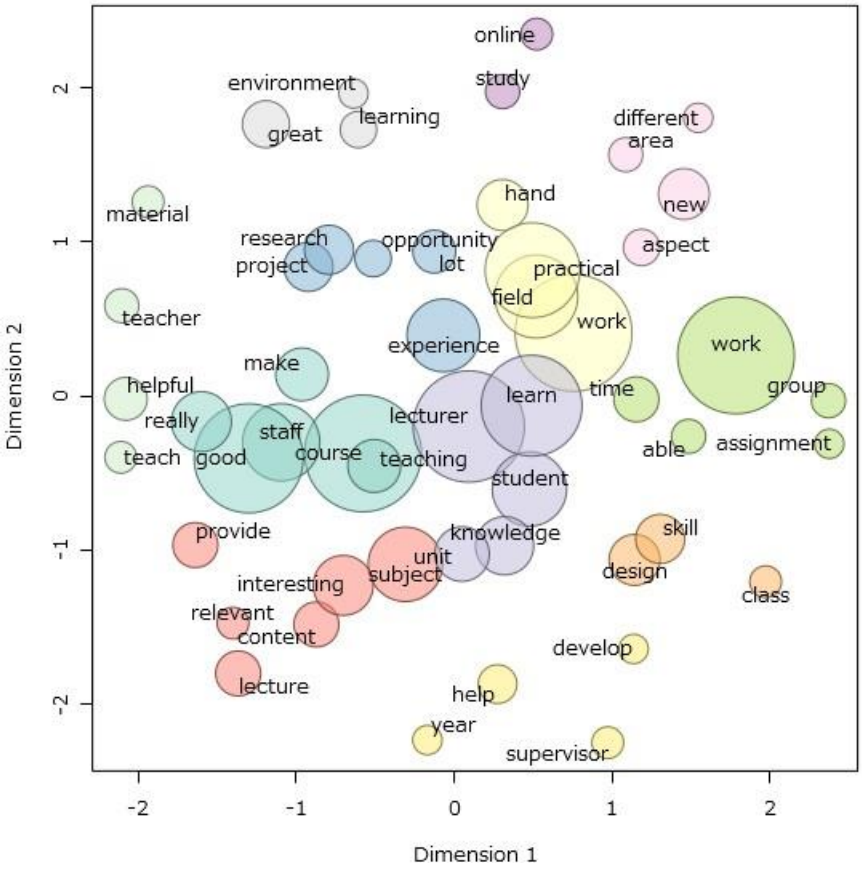

Figure 1: MDS visualisation for 'best aspects' comments for the entire Faculty

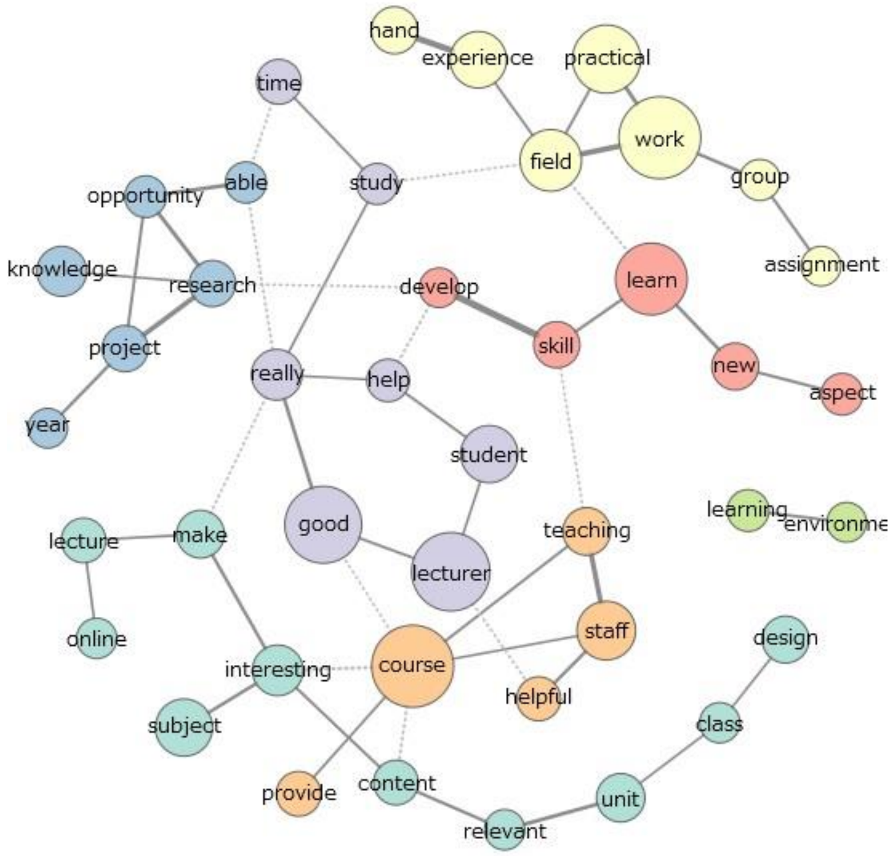

Figure 2: CON visualisation for 'best aspects' comments for the entire Faculty

Table 1: KWIC concordance for terms 'practical' & 'work' in BA comments for entire Faculty

	practical	**work** was very affective and applicable
the	**practical**	**work**
	practical	**work** was excellent
	practicals	and lab **work**.
	practical	**work**
able to apply knowledge acquired from theory and	**practical**	**work** to real life situations and able to…
…meeting new people and lecturers.	**practical**	**work** was good too.
use of online technology , interesting	**practical**	**work** ability to choose electives…
	practical	applications of field **work**
field, camps and	**practical**	**work**.
good	**practical**	field **work**
	practical	field **work** was thoroughly enjoyable…
	practical	**work** but the course could use some more
	practical	**work** and professional practice placement
the	**practical**	**work** that i did.
	practical	**work**
units where the	**practical**	and tutorial classes involved **working**…

Intra-Faculty/inter-School level

Following tagging of individual student comments with a School identifier (SCA, SCB, SCC or SCD), Figure 3 presents the MDS visualisation for BA comments for the entire Faculty. This visualisation is based on an analysis including terms with a frequency of 14 or greater, resulting in 48 terms being included. Figure 4 presents the CON visualisation for School-tagged BA comments for the entire Faculty based on the same set of 48 terms. While many of the same terms appear in Figure 3 and Figure 4 compared to Figure 1 and Figure 2, the slightly smaller number of included terms, and the appearance of the frequent School tags, means that some less frequently occurring terms have been omitted from these analyses. The relative size of the School identifier terms provides an indication of the relative number of BA comments received for each School. The presence of the School tags in the analysis means that the relationships between comment terms has been altered somewhat, with the School names acting a focus points ‘attracting’ those terms most frequently appearing in student comments associated with those Schools. It can be seen that School A (a design-based discipline School) is particularly associated with the term ‘design’, and interrogation of the term ‘design’ with the KWIC concordance tool revealed that virtually all comments including the term ‘design’ were from School A. School B appears in Figure 3 as a relatively small MDS bubble, but doesn’t appear in Figure 4 (CON) at all. The small size of the School B bubble in the MDS and its absence from the CON suggested that the comparatively few BA comments received for School B did not contain specific terms that occurred frequently enough to reach the threshold limit for inclusion in the CON visualisation. It can be seen that School C (a School hosting significant laboratory and field work) was strongly associated with the ‘practical work’ dyad (term pair) observed in Figure 1 and Figure 2. Interrogation using the KWIC concordance tool confirmed that this was the case. Figure 3 shows that School D appeared to be associated with the adverb term ‘really’. Figure 4 suggests that this could be in conjunction with the term ‘good’. Interrogation using the KWIC concordance tool confirms a number of student BA comments from School D contained the dyad ‘really good’. The intra-Faculty visualisation process was repeated for the NI comment set. Together, this set of four visualisations provides an additional Faculty-level overview of the key comment themes reported by students, including inter-School information about the relative number of CEQ comments from each School, and the relative association of each School with comment terms within the resultant 2D space of the each type of visualisation.

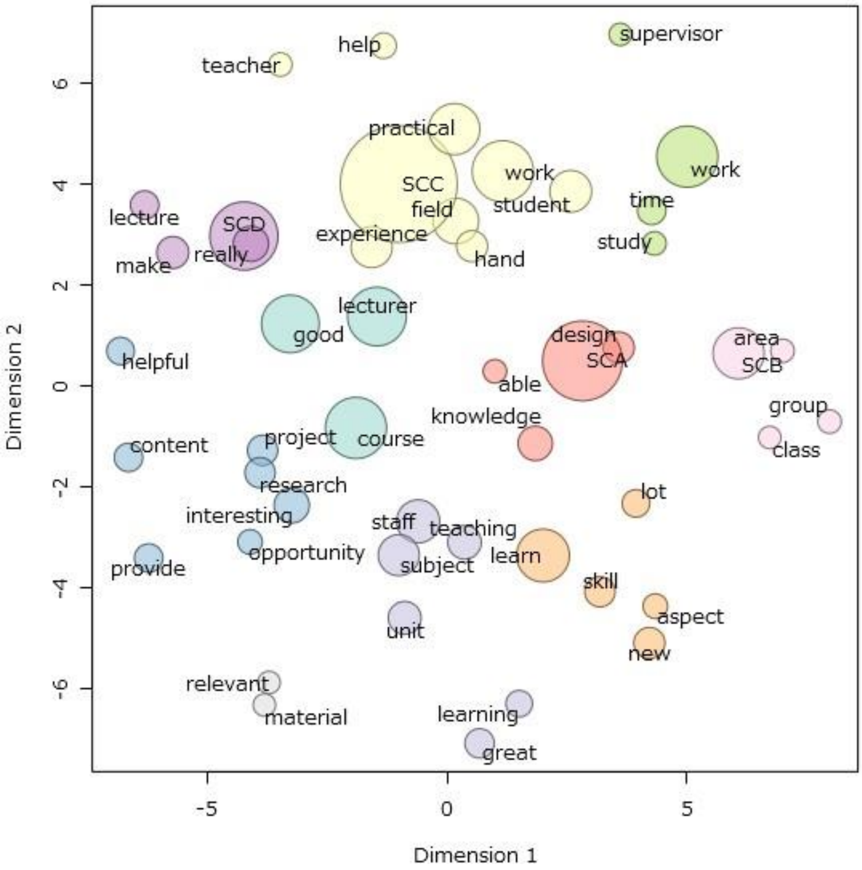

Figure 3: MDS visualisation for ‘best aspects’ comments for the intra-Faculty level

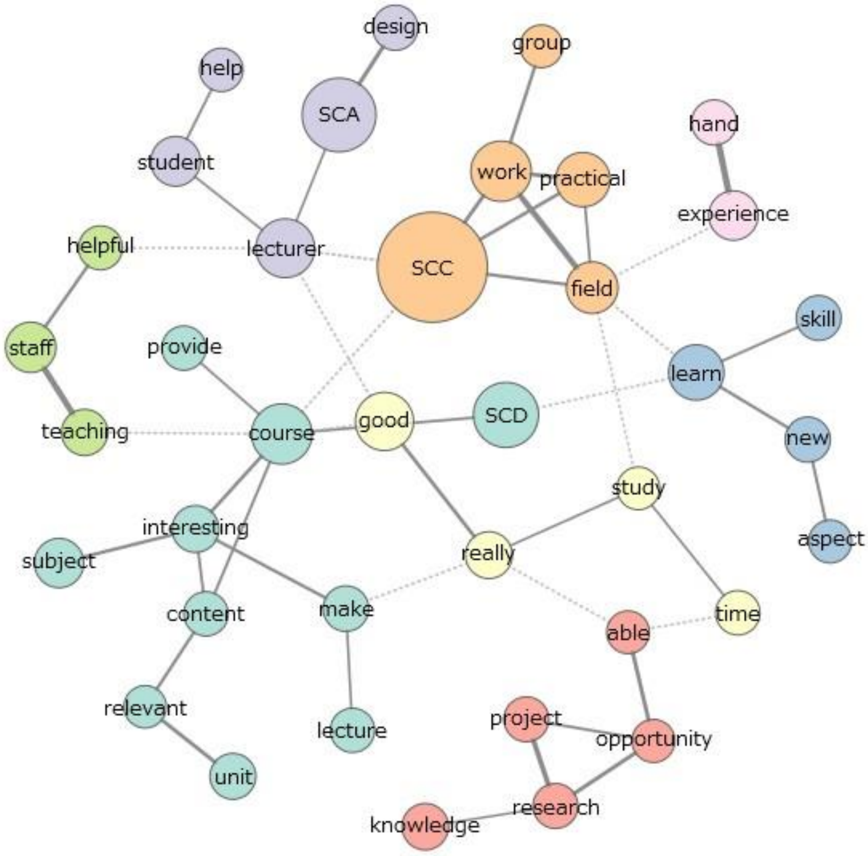

Figure 4: CON visualisation for 'best aspects' comments for the intra-Faculty level

School level

Although the details are omitted here for brevity, the student comment sets for each individual School were separately visualised using the Faculty-level method described above, to obtain a view of the comment themes specifically for each School.

Conclusions

A method for processing CEQ comment data and analysing them with the free KH Coder text analytics software package to produce relevant and informative visualisations was developed. Multi-dimensional scaling visualisations were found to provide a useful overall representation of the key words/themes in CEQ comment data, showing the relative relationship between words/themes. Co-occurrence network visualisations were found to provide a useful representation of the key word phrases/clusters in CEQ comment data. The KWIC text concordance feature allowed the comment data underlying the visualisations to be interrogated to understand the original context of the comments. Three different levels of analysis ((i) whole of Faculty; (ii) intra-Faculty/inter-School; and (iii) individual School) provided information yielding different insights into the student comment data for different levels of academic administration and leadership within the SEBE Faculty. In particular, the intra-Faculty level visualisations successfully identified some of the distinctive characteristics of particular Schools, such as a design focus and significant use of practical work. Although omitted for brevity, the various NI comment visualisations successfully identified many of the issues commonly reported by students in CEQ comments as 'needing improvement', including access to resources, opportunities for work experience, better assignment feedback, and more time with teaching staff.

We note some limitations to this investigation. While text analytics visualisation techniques provide an objective and repeatable representation of open-ended student comment data, it is still a manual task to interpret the results of the visualisations and take any action in response (Zaitseva et al., 2013). The 'first rule' of advice from one of the developers of the CEQ was that CEQ data should not be considered in isolation from other sources of information, such as other student evaluation of teaching surveys, benchmarking with relevant university partners, surveys of employers and graduates, and advice from accreditation bodies (Ramsden, 2003).

The text analytics method developed for analysing CEQ open-ended comment data using the KH Coder software package produced useful comment text visualisations that, in turn, provided a valuable perspective on these comment data in a straightforward and timely manner. The method developed and documented here is a practical and useful approach to analysing/visualising CEQ open-ended comment data that could be applied by others with similar comment data sets.

References

Bolden, R., & Moscarola, J. (2000). Bridging the Quantitative-Qualitative Divide: The Lexical Approach to Textual Data Analysis. *Social Science Computer Review, 18*(4), 450-460.

Chen, L., & Buja, A. (2009). Local Multidimensional Scaling for Nonlinear Dimension Reduction, Graph Drawing, and Proximity Analysis. *Journal of the American Statistical Association, 104*(485), 209-219.

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience, 21*(11), 1129-1164.

Higuchi, K. (2014). KH Coder (Version 2.beta.32). Japan: Koichi Higuchi. Retrieved from http://khc.sourceforge.net/en/

Hu, X., & Liu, H. (2012). Text Analytics in Social Media. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 385-414): Springer US.

Ishii, N., Suzuki, Y., Fujii, T., & Fujiyoshi, H. (2013). Development and Evaluation of Question Templates for Text Mining. In F. L. Gaol (Ed.), *Recent Progress in Data Engineering and Internet Technology* (Vol. 156, pp. 469-474): Springer Berlin Heidelberg.

Jackson, K. M., & Trochim, W. M. K. (2002). Concept Mapping as an Alternative Approach for the Analysis of Open-Ended Survey Responses. *Organizational Research Methods, 5*(4), 307-336.

King, W. R. (2009). Text Analytics: Boon to Knowledge Management? *Information Systems Management, 26*(1), 87-87.

Minami, T., & Ohura, Y. (2013, 31 August-4 September). *Investigation of Students' Attitudes to Lectures with Text-Analysis of Questionnaires.* Paper presented at the IIAI International Conference on Advanced Applied Informatics, Los Alamitos, CA.

Namey, E., Guest, G., Thairu, L., & Johnson, L. (2007). Data reduction techniques for large qualitative data sets. In G. Guest & K. M. MacQueen (Eds.), *Handbook for team-based qualitative research* (pp. 137-162). Plymouth, UK: Altamira Press.

Ramsden, P. (2003, 11-13 June). *Student Surveys and Quality Assurance.* Paper presented at the Australian Universities Quality Forum 2003 - National Quality in a Global Context, Melbourne.

Richardson, A. (2003, 30 November-3 December). *Qualitative Analysis of Graduate Comments and the Development of Course Domains.* Paper presented at the International Education Research Conference AARE - NZARE, Auckland.

Zaitseva, E., Milsom, C., & Stewart, M. (2013). Connecting the dots: using concept maps for interpreting student satisfaction. *Quality in Higher Education, 19*(2), 225-247.

Copyright