Eureka prize assessment

Innovation

Why and how was project conceived?

As an outsider to the project, my knowledge of this aspect of the project is fairly sparse. My entrée to the project came as statistical consultant and research methodologist for the Faculty of Education, Griffith University. In that role, I quickly sensed the merits of this text analysis engine and circulated information about Leximancer to the extent that Griffith University users have taken up 24 of the 100 plus licences issued to date.

It is my understanding that this project was conceived with the intent of providing researchers with an automated tool that could investigate associational patterns in text with ease and with the kind of reliability and validity standard associated with numerical analyses. While quantitative researchers have long since been able to use largely automated tools to conduct numerical analyses (e.g., SPSS) the same has not been the case for text. In these terms, Leximancer aligns text and numerical analyses insomuch as it quantifies the vast amount of the material produced by human beings in the form of text. It does so via an automated procedure that considers all text elements, identifies a core set of important elements or concepts, and then maps the frequencies and association between these concepts. In doing so, this software efficiently quantifies the relational patterns inherent in qualitative data, and produces replicable and valid output based on what is actually present in the dataset.

To what extent is project novel?

To date, the analysis of natural language data sets has relied on using the garage floor or using semi-automated software where the analyst's search for or selection of terms is based on previous theorising. The drawback with this theory-based (top-down) approach is its alignment with the subjectivity (preconceptions & views) of the qualitative data analyst.

As described in a forthcoming Behavioral Research Methods article (A.E. Smith & M. S. Humphreys, 2005), Leximancer remedies the situation by using a combination of novel techniques to scan and cluster terms in natural language data sets based on what is actually present in the text (i.e., bottom-up approach). One feature of particular interest to the text analyst is the high level of automation. This automation includes being able to open text files in a variety of formats, analyse files utilising default settings, study the output as a two-dimensional map, and easily view text links within text. Alternatives to Leximancer cannot boast this combination of features.

Has the project developed or extended existing techniques or methods?

The notion of text analysis is not novel but Leximancer's adaptation of existing procedures fits it to be described as the first truly modern text analysis or text mining program. The remainder of this section outlines some starting points on which Leximancer has built or improved.

Thesaurus construction software has been around for a while, and theory based semi-automated text analysis software has enabled users to search text and obtain result summaries that include the frequencies of search terms. A subset of more modern text analysis programs have taken this process further by offering variants on the notion of automatically collecting terms and plotting the strength of association between those terms. However, on inspection, these programs are either at an early stage of development, can identify terms and associations but are limited in their capacity to represent the results other than within a moving text window, rely on extensive domain specific dictionaries culled from natural

language, or are poorly automated. In short, other text analysis programs perform similar tasks but fail the test of ease of use and usability.

Impact

In what ways will the completed project affect the subsequent study, science, and application of information and communication technology?

I'm not a programmer so cannot comment in any depth on this aspect of Leximancer's effect. However, I'd venture that insomuch as information technology has built on attempts by logical positivists and others to formalise the meaningfulness of surface language in terms of logical propositions, that Leximancer's ability to extract propositions directly from samples of natural language is evolutionary. That is, rather than relying on formalised and rule-based approaches to analysing natural language, information technologists now have the option of using the more flexible approach pioneered by Leximancer.

What are the broader technical implications of the project?

Again, I'd have difficulty in commenting on Leximancer's technical implications. But I'd venture that its advanced level of automation, robust nonparametric algorithms, and general ease of use combine to challenge the notion that qualitative materials cannot easily be analysed systematically. For this reason, I'd expect the presence of Leximancer to intensify the debate about the usefulness of qualitative data. Not only that, but as with any tool that offers the possibility of a paradigmatic shift, its ability to generate novel and replicable findings have the potential to revise judgments about what is worth observing. In short, I'd expect this project to impact on the study, science, and other aspects of information and communication technology strongly insomuch as it opens up a capacity to examine materials that to this point were largely unanalysable.

What are the broader (industry and/or community, social and/or economic) implications of the projects?

Thomas Kuhn (1977) wrote about "The essential tension" between various paradigms or approaches to doing science, and he could easily have been referring here to the tension between the practitioners of qualitative versus quantitative forms of data analysis. This tension has surfaced in endless debates about the imbalance in replicability, reliability, and validity offered by analytic procedures from these two domains, with qualitative analysis as the poor cousin. Now Leximancer offers qualitative investigators the option of conducting systematic analyses of text in naturally occurring forms, including emails, open-ended answers in questionnaires, the instruments themselves, the transcribed talk of focus groups, as well as material from journal articles, books, etc.

At my place of work, Leximancer is being used to re-examine the important associations implicit in text-based data on topics related to emotional intelligence, government policy, conference proceedings, legal materials, marketing materials, and multimedia, in other words, a broad range of text based materials. I'd expect Leximancer's usefulness to extend from academic settings through commercial interest and into policy considerations. More specifically, Leximancer could be expected to impact areas as varied as market research strategies and national security.