Deze rubriek is een formum voor debat over kwalitatief methodologische kwesties. Wie een idee heeft voor een thema of wil reageren op een eerder verschenen bijdrage, vragen wij contact op te nemen met Fred Wester: F. Wester@maw.ru.nl.

Using Computational Techniques to Fill the Gap between Qualitative Data Analysis and Text Analytics

Curtis Atkisson, Colin Monaghan and Edward Brent

Introduction[1]

The recent mass digitization of text data has led to a need to efficiently and effectively deal with the mountain of textual data that is generated. Digitized text is increasingly in the form of digitized data flows (Brent, 2008). Digitized data flows are non-static streams of generated content – including twitter, electronic news, etc. An oft-cited statistic is that currently 85% of all business data is in the form of text (cited in Hotho, Nürnberger & Paass, 2005). This mountain of data leads us to the question whether the labor-intensive traditional qualitative data analysis techniques are best suited for this large amount of data. Other techniques for dealing with large amounts of data may also be found wanting because those techniques remove the researcher from an immersion in the data. Both dealing with large amounts of data and allowing immersion in data are clearly desired features of any text analysis system.

The three major extant solutions to this problem each address important problems, but have some issues. An initial description of each of the three solutions is presented.This is followed by the presentation of the methodology and motivation behind a recently developed system, Veyor®. This system will then be used to analyze the dataset on the economic downturn created by KWALON. Results will be discussed along with program usability and user experience.

Existing methodologies

The three most frequently applied methodological solutions to analyzing qualitative data are qualitative data analysis (QDA), content analysis and text mining. Each of these methodologies offers unique ways of approaching data and reporting results.

1 The authors would like to acknowledge Luis Cruz and Adam Khalil for coding assistance. We would also like to thank Matt Wood, Nate Green, Brian Cooksey and James Barton – members of the Veyor development team.

Qualitative Data Analysis (QDA)

QDA emphasizes immersion in the data and an open coding procedure (e.g., Glaser & Strauss, 1967). This procedure can be pursued either by eschewing any preconceived notions, or by making those preconceived notions explicit prior to analysis. This methodology, like all methodologies, has a series of strengths and weaknesses.

The key strength of qualitative data analysis lies in the stated aim of allowing the data to speak for itself. This places an emphasis on the actual statements made by various participants in the study, allowing for previously unconsidered connections to be elucidated. This methodology also places an emphasis on interesting statements above the most frequently mentioned causes or connections, typically not reporting any comparative statistics.

QDA has three primary weaknesses: time required, reliability and generalizability. These are particularly acute for digitized data flows. The time required for multiple readings of the data make even a traditional QDA study a daunting task, and the added time can lead to added costs. A traditional, open coding QDA study encourages the organic construction of a coding scheme, which has lead people to question the reliability of the results (Morse, Barrett, Mayan, Olson & Spiers, 2002). While this has partially been addressed through inter- and intra-rater reliability, the procedures employed often emphasize assessing reliability while doing little to ensure reliability, only assess reliability for a small subsample of data, and add substantially to the time and cost of the study. Generalizability is limited by both the appropriateness of generalizing a coding scheme to another study and the additional time required to code a new study using the same framework. The costs of generalization discourage reanalysis and replication studies (Thompson, 2000).

Content Analysis

Content analysis emphasizes producing results that are reliable and generalizable (Holsti, 1969; Stemler, 2001), and typically relies little on advanced computational techniques. Content analysis often uses simple word frequency and key-word-in-context statistics to elucidate patterns in data. While very fast, these statistics may miss important insights and have only limited internal validity. Due to a desire to achieve reliability, content analysis often places less emphasis on emergent themes derived from an in-depth examination of the data. Generalizability is achieved by making the coding rules as explicit as possible (Mayring, 2000).

Hsieh and Shannon (2005) distinguish three distinct variants of content analysis – each with their own strengths and weaknesses.

Conventional content analysis is descriptive and avoids using preconceived categories (Kondracki, Wellman & Amundsonc, 2002). This has the benefit of immersion in the data, the use of inductive category development (Mayring, 2000) and the goal of reducing the number of code clusters to allow for interpretability (Morse & Field, 1995). Negative considerations include not achieving internal validity, i.e. not getting correct context for codes; often countered through prolonged exposure to the data (Lincoln & Guba, 1985).

Directed content analysis uses an existing theory or body of research to direct further examination into text data. Using deductive category application (Mayring, 2000), this methodology allows one to use rank order statistics to make statements regarding frequency of code application (e.g., Curtis et al., 2001) and allows for the coding of emergent themes not present in the initial theory. This may pose an issue to the naturalistic paradigm given that the initial code scheme may bias code application (Hsieh & Shannon, 2005).

As the most quantitative of the strategies, summative content analysis uses quantification of words, or other meaningful units, to explore the usage of concepts. This includes both manifest coding, the actual words on the page, and latent coding, the underlying meaning (Babbie, 1992). Of the strategies, this is most dependent on the validity statistics previously mentioned. This strategy most immediately allows for generalizability, but may cause issues such as bias in the initial code application.

Text Mining

Text mining uses the power of modern computational technology. This technique seeks to automatically identify themes and trends in qualitative data. This technique seeks to identify patterns of interest and apply those patterns to different areas. The newest of the three approaches, text mining has its own set of weaknesses and strengths.

The key benefits to text mining are reliability and time savings. Because these programs use a computer to automatically code, code application is 100% reliable. The coded text from a text mining analysis can then be examined in a similar fashion to traditional qualitative analyses. Because text mining can code so much text so rapidly it can also generate enough data to justify the use of quantitative analysis techniques, including measures of association, tests of significance, and regression. Furthermore, using automatic pattern recognition and code application can result in drastic time savings – allowing us to deal with digitized data flows (Brent, 2008).

The primary concern of qualitative researchers regarding text mining is the reduced role of the researcher in analyzing the data. The premium placed on minimizing human effort in text mining raises concerns about the internal validity of the results (as discussed in reference to content analysis) and the ability to identify meaningful emergent themes.

Filling the Gap Between Methods

As is evident from the above discussion, each of the three major approaches to the analysis of qualitative data has strengths and weaknesses. QDA provides the researcher with the greatest control over the project, but is the most time intensive. Content analysis has the ability to use frequency of occurrence statistics and is valid and reproducible, but it can take as much as or more time than QDA and it employs relatively unintelligent computational strategies. Text mining places such a premium on saving time that the input of the researcher can be lost.

Each of the three traditions is strongest when human researchers and the computer are assigned to tasks for which they are best; they are weakest when tasks are assigned to a computer that would have better been done by a human or vice versa. For instance, humans are strong at recognizing when a code applies to a segment of text, but weak when it comes to performing this highly repetitive task consistently for large amounts of text. Computers are initially bad at recognizing when a code applies to a segment of text, but once trained to do so accurately, can perform the task quickly and reliably for large volumes of data. QDA, even computer-assisted qualitative data analysis systems (CAQDAS), are strong on insight and validity but weak on reliability and speed because computers only assist while humans carry out the coding and analysis. Data mining and content analysis reverse the emphasis having computers perform most of the work with only light human input, and hence they are weaker on validity and insight but stronger on reliability and speed.

Here we briefly describe a new program, Veyor®, and indicate how it draws from these three traditions in an effort to exploit the benefits from each while avoiding their limitations. Given the limitations of space, a more detailed description must be presented

Figure 1. Specifying relationships between codes

elsewhere. Veyor uses a hybrid strategy combining the human coder with a partially automated coding process, applying each to what they do best and reducing the cognitive load on the researcher. With Veyor, as in QDA, the researcher plays a strong role in those tasks at which humans excel: identifying emergent themes and developing an evolving coding scheme. Unlike QDA however, with Veyor the researcher only codes a sample of cases from the data. Those coded cases are used to train the computer program to code accurately and then the program is used to assign codes to the great majority of the data. In addition to the correct apportioning of effort, we sought to decrease cognitive load for the researcher at all times (Nielsen, 1993). As such, at any one time in the process of interacting with Veyor, the user is only asked to perform a single task.
First the researcher examines a sample of data to identify emergent themes or codes. Then the computer applies those procedures to automatically code the entire data set. Once all segments are coded by Veyor, the researcher validates the coding by assigning their own codes to a random sample of data segments then comparing their codes with those of the computer. The process is repeated for one code at a time until the coding of the dataset has been validiated to a pre-specified level. The program can then be used to automatically generate any of a series of standard or user-specified reports including frequency charts, cross tabulations, pareto analyses and segments for each code.

This approach has a number of strengths. It can be employed to code very large datasets, including digitized data streams of massive amounts of data, producing summary statistics that facilitate a quantitative analysis. Yet it is also carefully validated to meet a user-specified standard of validity, employs a coding scheme carefully developed and validated by the researcher to reflect human insights or to focus on concepts relevant to a particular theoretical perspective, and can be used to conduct a qualitative analysis much like that produced by the QDA approach.
There are, however, a number of potential weaknesses of Veyor. Those weaknesses are not inherent in Veyor but reflect weaknesses that would result from misuse of the program and can be minimized or eliminated with effective planning and protocols. The ease of automated coding could discourage a researcher from full immersion in the data. The validity standard used is set by the researcher and too conservative a standard may leave researchers spending more time than strictly necessary on validation. Without careful effort paid to the open-coding aspect of the program, including examining random samples of segments, the study may lack validity. In order to put this program in context we illustrate its use to analyze the same dataset analyzed by several other qualitative analysis programs at the recent KWALON conference.

Current Study and Methods

The current study examines data on the economic crisis collected by KWALON for use in their conference on qualitative data analysis systems. The final dataset contained 97 newspaper articles and 102 blog entries totaling over 5800 sentences. We analyzed this dataset to investigate worldwide views of the economic collapse as revealed in traditional and

social media sources. Specifically, this study asks the question: *'What are the primary actors, causes and consequences mentioned in relation to the economic crisis?'*

To frame our research inquiry, three broad coding categories were established in Veyor before beginning analysis: (1) actors; (2) causes; (3) consequences. Next, two trained qualitative data coders manually analyzed a random subset of the data. Working with sentence-sized coding units, coders viewed a selection of segments one at a time to establish new codes, placing them into one of the three broad code categories. Existing codes were expanded to include additional features found in the data.

Instead of coding segments directly, coders identified meaningful and distinguishing features required to recognize when a code should be applied to a segment. This information, paired with a number of built-in computational intelligence strategies, allowed Veyor to automatically assign codes in the background. In this way, the coders were able to train the program to recognize and differentiate significant concepts, while leaving much of the tedious work to Veyor's automated coding processes. The emergent coding phase concluded when coders reviewed 25% of the data which took about 40 hours.

By assigning codes automatically, Veyor achieves 100% reliability. To measure validity, we measured the level of agreement between the program and human coders. For each code, the program generated a random sample of segments and a human coder indicated whether or not each segment was correctly coded with the selected code. If the resulting comparison did not meet the desired standard, the rules used to automatically apply the selected code were refined to correctly code the sample then a new independent test sample was selected. Any changes to the rules for coding automatically recoded all segments. We required agreement of 90% accuracy with 95% confidence. The process continued until program performance exceeded the desired standard or all relevant segments had been confirmed by a human coder. Depending on the code's level of complexity, validation took between 5 and 25 minutes per code with the majority of codes requiring less than 10 minutes per code and a total time of roughly 40 hours.

Results and Discussion

The final coding scheme contained 239 codes, including 74 *actors*, 17 *causes* and 148 *consequences*. For clarity, an additional level of organization was added to the actors and consequences categories, creating a three-level hierarchy. For this study we employed a mixed-mode (qualitative and quantitative) analysis using standard reports in Veyor to summarize the relative frequency of occurrence of common codes along with segments to illustrate those codes. All the codes for a segment are given directly below and to the right of that segment.

Pareto analysis per category with sample statements

This Pareto analysis emphasizes the most commonly occurring categories and helps us compare their relative frequencies. A single segment can be counted in multiple categories.

Actors

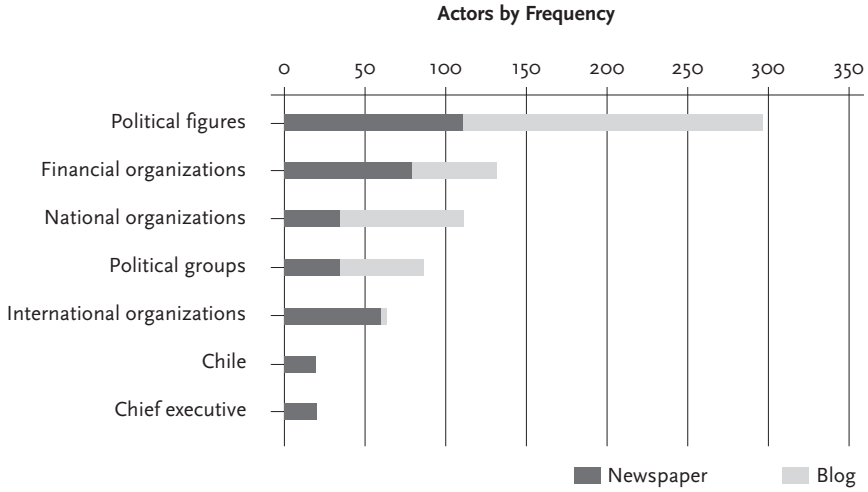

Figure 2

The most common *actors* identified in the data were political figures (N=297), making up over 40% of the actors category. Most political figures were from the United States, like Barack Obama (N=63), Ben Bernanke (N=47), and George W. Bush (N=44). Other notable politicians include Russian President Dmitry Medvedev and German Chancellor Angela Merkel.

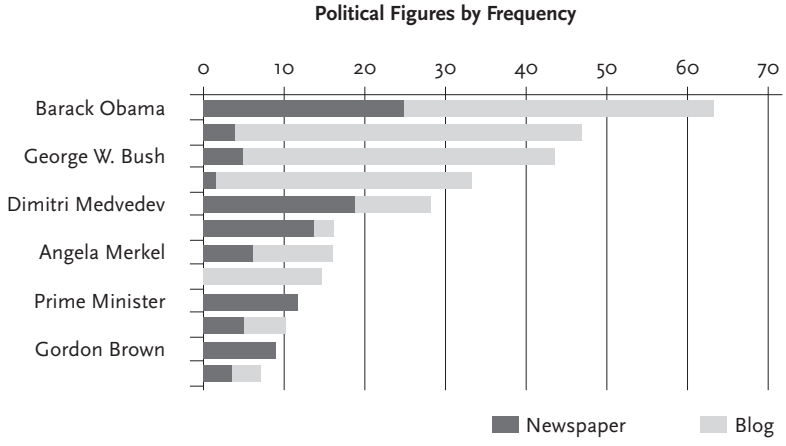

Figure 3

Other frequently mentioned actors include financial organizations (N=131), national organizations (N=113) and political groups (N=92).

Actor sample statements

> The Obama administration appears to be using the current economic crisis to create a bigger role for government throughout the economy, from education and health care to banking and energy, in a bid to get the crisis under control.
>
> *Big government; Obama*

> Financial stocks have gyrated in response to the Obama administration's plans, which promise tighter oversight and new, tougher rules for banks.
>
> *Obama*

> 'A lot of things happened, a lot came together, and created probably the worst financial crisis, certainly since the Great Depression and possibly even including the Great Depression', Bernanke said at the start of a town-hall meeting in Kansas City.
>
> *Bernanke*

> It was certainly thought, by Bushy and friends, that the corporate war profits would at least bring prosperity to a few, it did to very few but the downturn in oil they were hoping for backfired and yes they are responsible for the rest of it.
>
> *George W. Bush*

Causes

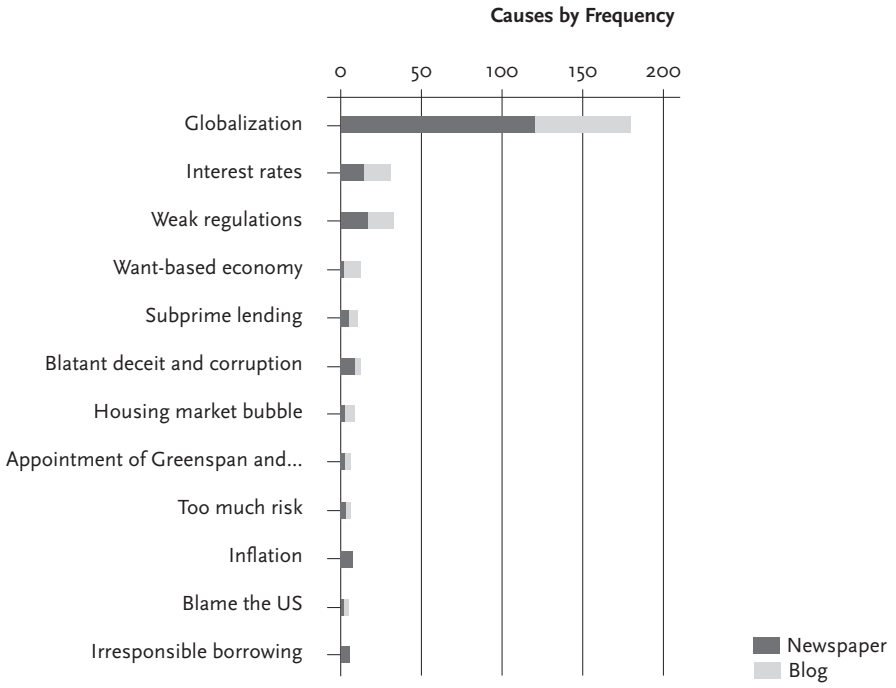

Figure 4

Three *causes* in particular stood out as the most commonly occurring. Globalization (N=177) was the most frequently mentioned cause, followed by issues with interest rates (N=31) and weak regulations (N=30).

Causes sample statements

> In its annual report on Monday it called for an overhaul of financial regulations, economic policy and the structure of the global economy.
>
> *Globalization*

> The manufacturing sector in the United States has been destroyed by globalization.
>
> *Globalization*

> Without the forces of globalization, at least globalized capital, then the financial institutions would not have been able to borrow, lend and collateralize then borrow, lend and collateralize again.
>
> *Globalization*

> The Fed's easy money policy is now stoking US inflation rather than a recovery.
>
> *Low interest rates*

> The cheap dollars that the Feds are printing is causing a monumental inflation in commodities such as food and energy, which is causing a world crisis.
>
> *Low interest rates*

> 'These results show a public sobered by a financial crisis precipitated by weak regulations and a lack of corporate accountability', said Transparency International chairperson, Huguette Labelle.
>
> *Weak regulations*

> Some boards have not only failed in the oversight of risk management systems but also in the remuneration practices of their firms, so the financial market collapse was their failure, too.
>
> *Weak regulations*

Consequences

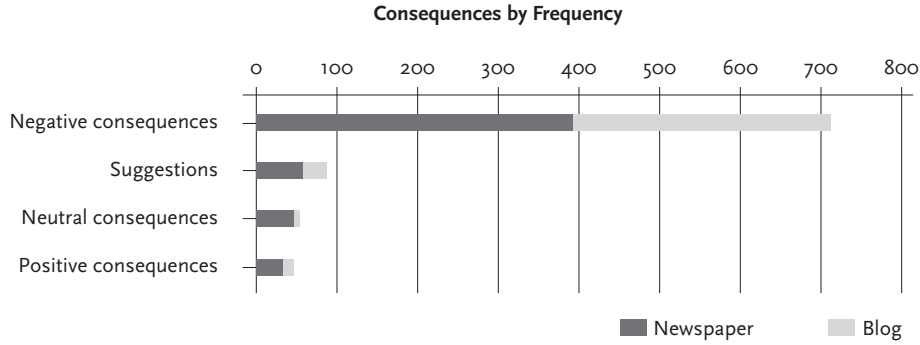

Figure 5

Not surprisingly, negative *consequences* (N=719) of the economic crisis far outweighed positive or neutral consequences.

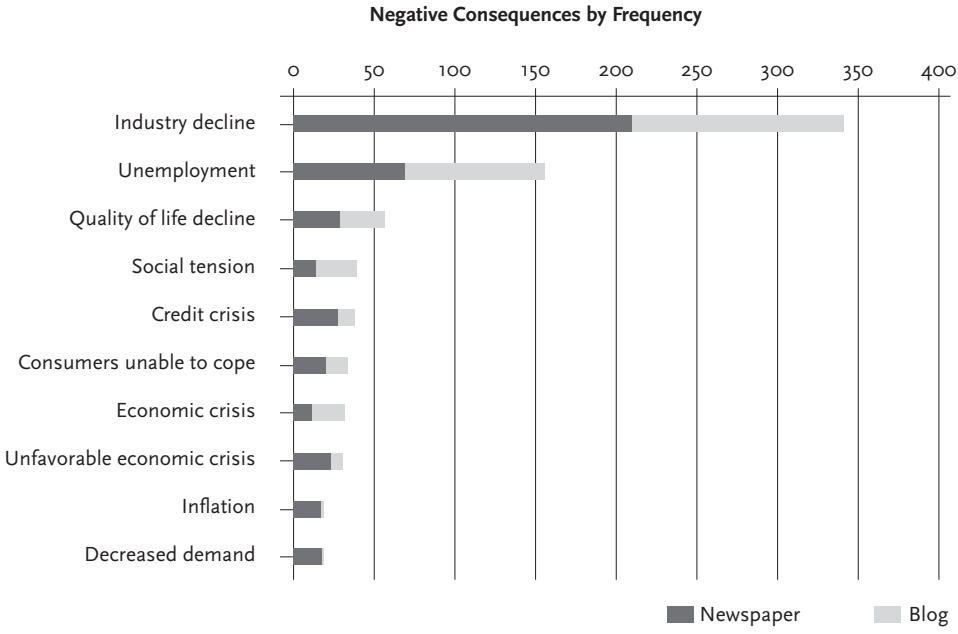

Figure 6

Top negative consequences include industry decline (N=341), unemployment (N=154) and a decline in the quality of life (N=56).

Consequences sample statements

> Investors are worried about inflation and central bank rate rises because prospects for global growth and company earnings look bleak.
>
> *Low profitability; Globalization; Inflation*

> Recession hits airlines especially hard Albert Jung, a corporate communications manager at IATA, says the airline industry has been one of the hardest hit by the global economic meltdown.
>
> *International Air Transport Association; Hardest hit; Globalization*

> Another area of concern for both central banks is a possible slowdown in the property market, where banks are exposed via loans to commercial property companies and mortgages.
>
> *Slowdown in the property market*

> Some economists claim that national unemployment could surpass 12 percent during the winter months (June through August), when seasonal jobs in agriculture and construction are in lower demand.
>
> *Higher unemployment*

> Millions of people will lose everything they've got and won't be able to find a job for a few years...
>
> *Higher unemployment*

> The cost of daily living, from buying food to getting medical care, will become difficult for all but a few as the dollar plunges.
>
> *Cost of living*

Newspaper and blog comparison

Data analyzed here were taken from both blogs and newspapers, raising a question as to whether there are important differences in comments found in those two sources. Blogs and newspapers are not significantly different in the relative frequency of the top ten causes ($\chi^2 = 15.4$, df = 0, p = .081). But they are significantly different in both actors and consequences. Blogs are more likely to mention political figures, political groups, and national organizations as actors; while newspapers devote more attention to financial organizations and international organizations ($\chi^2 = 141.2$, df = 6, p < .000).

While both newspapers and blogs both emphasize negative consequences, blogs are relatively more likely to focus on negative consequences, and newspapers are more likely to focus on neutral or positive consequences and suggestions ($\chi^2 = 21.6$, df = 3, p < .000). Comparing the negative consequences for blogs with those for newspapers there is a very strong pattern, with newspapers devoting more attention to macro-level economic issues such as industry decline, credit crisis, inflation, and decreased demand; and blogs

focusing more on individual-level consequences such as unemployment, social tension, and consumers unable to cope. The different coverage is statistically significant (χ^2 = 41.8, df = 9, p < .000).

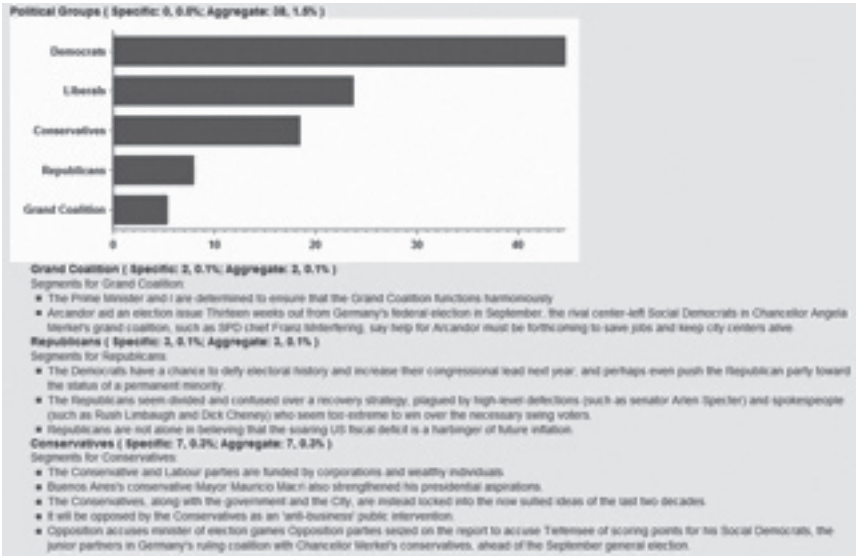

Figure 7. One type of report from Veyor

Discussion of results

A closer look at the data can help explain why globalization is by far the most frequently mentioned cause. First, it may indicate unfamiliarity with more technical explanations. Individuals without an awareness of advanced economic principles might be unable to name precise causes, like subprime mortgages or Alan Greenspan's policies, and instead mention a more straightforward trigger like globalization. Second, while other incidents may have caused regional problems, there appears to be agreement that globalization is the reason economic disturbances spread worldwide.

Also, it's worth exploring the relationship between causes and consequences. Because of the complex relationships between economic activities, the line between causes and consequences is often blurred. For example, interest rates emerged as the second most common cause. At the same time, many sources mentioned the fact that interest rates are being lowered around the world to help jumpstart the economy. This is a clear consequence of our economic situation, but may also be an ongoing cause if lower interest rates aggravate the recession.

User experience

This study demonstrates the benefits of a hybrid approach to qualitative analysis. By defining and distinguishing significant concepts found in the data and allowing the program to automatically assign codes, coders were able to complete a thorough analysis

while only viewing 25% of the data. This process decreases manual effort and encourages an adequate level of data immersion without becoming impractical for larger datasets.
Another benefit of building a coding scheme in Veyor is that new data can be added to the program and coded automatically. This dramatically reduces personnel costs and boosts coding speed for subsequent datasets, making Veyor cost-efficient for large or ongoing sets of data. In addition, the trained and validated coding scheme guarantees coding consistency across projects so that future studies can be compared directly. This system could be combined with other methods, with Veyor providing ongoing monitoring and some level of analysis, supplemented – when deemed appropriate – by further analysis.
Coders reported taking advantage of Veyor's two-step process to optimize speed and accuracy. During the coding phase, codes were intentionally defined broadly to ensure the program didn't omit relevant segments. Then by examining a random sample of segments during the validation phase, coders improved validity by refining each code and reducing false positives. This approach allowed coders to focus on one objective at a time and maximized the efficiency of human effort. It also provided a solution to the age-old problem in QDA of whether to recode data when new codes emerge. With Veyor every time a code is added or modified the program automatically recodes all the data in the background. Veyor also accommodated multiple coders without difficulty. Users easily modified and improved their shared coding scheme, without needing to coordinate recoding work. Because Veyor utilizes a database accessible over the Internet it was even possible for coders to work simultaneously on the project from different geographic locations.
The coders identified one issue that could increase efficiency in the future. In this study, the coding unit was defined on the sentence level. This occasionally prevented the user from understanding the context of the segment when presented for coding or validation. It can be difficult to discriminate the features of an isolated segment. Subsequent versions of Veyor could resolve this issue by presenting segments in context for coding.

Conclusion

As discussed in the introduction, a system that is well-designed to address the analysis of our increasingly large digital data flows will allow a researcher to obtain immersion with the data, but not result in exorbitant costs of time or money. A candidate system, Veyor, was presented which seeks to merge the strengths from the three traditions of QDA, content analysis and text mining while avoiding their weaknesses. This system was used to examine a dataset examining a large number of sources speaking about the recent economic recession and found to offer several advantages. As with any system having a human component, this system also has several potential weaknesses that can be reduced through established procedures and protocols.
This system appears to be a candidate for wide application in both industry and academia. By assigning tasks appropriately and reducing human cognitive load, Veyor allowed researchers to build an original coding scheme based on a large dataset with minimal

manual effort that could also be used to code data from the same ongoing data stream or completely different projects.

References

Babbie, E. (1992). *The Practice of Social Research*. Belmont, CA: Wadsworth.

Brent, E. (2008). Artificial Intelligence and the Internet. In N.G. Fielding, R.M. Lee & G. Blank (Eds.), *The Handbook of Online Research Methods*. London: Sage.

Curtis, J.R., Wenrich, M.D., Carline, J.D., Shannon, S.E., Ambrozy, D.M. & Ramsey, P.G. (2001). Understanding physicians' skills at providing end-of-life care: Perspectives of patients, families and health care workers. *Journal of General Internal Medicine, 16*, 41-49.

Glaser, B.G. & Strauss, A.L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine.

Holsti, O.R. (1969). *Content Analysis for the Social Sciences and Humanities*. Addison Wesley.

Hotho, A., Nürnberger, A. & Paass, G. (2005). A Brief Survey of Text Mining. *LDV Forum, 20*(1), 19-62.

Hsieh, H.-F. & Shannon, S.E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research, 15*(1277).

Kondracki, N.L., Wellman, N.S. & Amundson, D.R. (2002). Content Analysis: Review of Methods and Their Applications in Nutrition Education. *Journal of Nutrition Education and Behavior, 34*(4), 224-230.

Lincoln, Y.S. & Guba, E.G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Mayring, P. (2000). Qualitative Content Analysis. *Forum: Qualitative Social Research, 1*(2).

Morse, J. M., Barrett, M., Mayan, M., Olson, K. & Spiers, J. (2002). Verification Strategies for Establishing Reliability and Validity in Qualitative Research. *International Journal of Qualitative Methods, 1*(2).

Morse, J.M. & Field, P.A. (1995). *Qualitative research methods for health professionals (2nd Ed)*. Thousand Oaks, CA: Sage.

Nielsen, J. (1993). *Usability Engineering*, Morgan Kaufmann.

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation, 7*(17).

Thompson, P. (2000). Re-using Qualitative Research Data: A Personal Account. *Forum: Qualitative Social Research, 1*(3).