Using Automated Text Analysis to Study Self-Presentation Strategies

Eleanor T. Lewis, Jana Diesner, and Kathleen M. Carley

July 12, 2001

Extracting and representing the networks of ties between concepts in a set of texts creates a "map" of each text. Map analysis allows a researcher to compare the networks of ties between concepts in these texts by systematically reducing their content. The goals of this research paper are to answer both a methodological and a substantive question. First, how do the choices a researcher makes about how to generate maps using an automated text program alter the results, and how do these results compare to the results of hand-coding? Second, how can we interpret the results of map analysis to better understand the strategies authors use to manage their self-presentation, a central purpose of many texts. The texts we use are a subsample of a dataset of applications by entrepreneurs for an "Entrepreneur of the Year" award. Applicants value uniqueness in their application's *content* because it sets them apart and demonstrates their worthiness for the award, but the value placed on uniqueness in the *structure* of their accounts is not as clear. Our analysis allows us to extract four general self-presentation strategies: the prepared entrepreneur, the driven entrepreneur, the creative niche entrepreneur, and the humble entrepreneur (a single entrepreneur may employ multiple strategies).

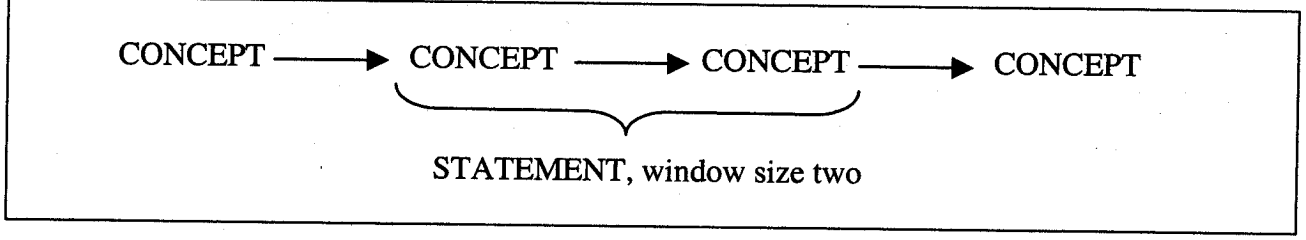
Text analysis or content analysis is a general term that describes "*any methodical measurement applied to text (or other symbolic material) for social science purposes*" (Shapiro & Markoff, 1998, 14; emphasis in original). Previous content analysis research has focused on establishing the validity of the technique for drawing inferences about texts, and on the challenges of using quantitative analyses to reduce the complex information in texts (for examples see Roberts, 1997). There has been less focus on any implications for interpreting the results of these quantitative analyses when the analyst makes different choices about *how* to analyze the texts (Carley, 1997a).[1] This issue is particularly important if an analyst uses map analysis to analyze concept networks because the ties between concepts are not explicit (like the ties between people—social networks), but implicit, linked grammatically and semantically through the sentences and paragraphs of the text. In map analysis, a *concept* is a single idea or ideational kernel represented by a single word or phrase, and a *statement* is two concepts and the relation between them; the map is the network of concepts formed from statements.

A text analyst can choose from a variety of data reduction techniques to extract the implicit ties between concepts in a text (Carley, 1993; Carley 1997a), including eliminating specific types of words (i.e. prepositions or verbs; i.e. Corman et al, 2001) and translating specific words into more basic concepts (Carley, 1997b). Both of these techniques are ways to reduce the text to more basic words or concepts that capture the features of the text a researcher is interested in. In addition, a researcher using map analysis must choose how to define what constitutes a tie between concepts. One technique is "windowing," a proximity based approach to establishing ties (i.e. Danowski, 1993). The analyst sets the size of the window, and concepts within that window are then "tied" to each other. For example, with a window size of two, every pair of words or concepts within two units of each other are linked in the text's network (see Figure 1). In this paper, we study the results of applying these data reduction techniques and varying the operationalization of window size.

---

[1] Carley (1997a) strongly recommends returning to the data several times and analyzing them under different coding conditions.

Figure 1. Visual representation of window size two



We analyze a subsample of 40 texts from a dataset of 174 applications from entrepreneurs to a foundation for an entrepreneurship award. The entrepreneurs operate in four primary industries: media, telecommunications, tele-electronics, and computer hardware and software. A typical application has several sections, beginning with a profile of the entrepreneur (used for the analyses here), followed by information about the company, their innovative practices, and their plans for the future of the company. The AutoMap© software program performs the data reduction and creates concept maps using different window sizes. AutoMap© contains map and content analytic routines that extract and analyze the maps of ties between concepts in texts.

AutoMap© allows the researcher to use a two step process for data reduction: deletion and generalization, both mentioned above. Deletion removes words from the text which do not help answer the research question (e.g. proper names, articles, and prepositions). AutoMap© has two *delete lists* available—an extensive one and a limited one—and the researcher can modify these or design a unique one. Generalization involves the application of a *thesaurus*, which must be designed specifically for a dataset. AutoMap© uses the entries in the thesaurus to search the text and "translate" specific words and phrases into more basic concepts specified by the researcher. In this dataset, the first author created the thesaurus through an iterative process. She read hundreds of previously published entrepreneur profiles, identified concepts that appeared repeatedly in these profiles, then checked them against a thesaurus and the words and phrases in a random subsample of the actual dataset (we will discuss these general concepts later). In addition to just applying a thesaurus, a text can be reduced to *only* the concepts in the thesaurus (the program replaces other words with symbols). AutoMap© provides statistical outputs for data analysis and visualization, and every step of data reduction can be stored for further analysis.

We chose to examine the results of three different approaches to data reduction. The first approach is hand coding. The first author coded each of the 40 texts using the concepts in the thesaurus, creating statements where she saw concepts semantically linked. This coding theoretically represents a "best case scenario" for what a map analysis program could accomplish in terms of locating and linking concepts into semantically appropriate statements. The second and third approach both use AutoMap© for deletion and generalization, and to create the statements in each text's map using the concepts in the thesaurus. The second approach creates statements using a window size of three. The third approach creates statements using a window size of 19 (the average length of a sentence in the texts), but modifies the operationalization to be the location of the concepts in the *original* text. It is unclear based on previous research (i.e. Carley, 1993) how this modification of the definition of window size will influence the coding results. For example, does it represent greater or lower data reduction? Are the concepts more strongly or weakly linked?

A text in our dataset has an average of 335 words before any data reduction. After applying a delete list, that text is reduced to an average of 221 words, a 34% reduction from the

raw text. Finally, applying the thesaurus to the text after the delete list further reduces the text to an average of only 20 total concepts. This represents a 91% reduction from the text after the delete list, and a 96% reduction from the raw text. The highest percentage of a text captured by the concepts in the thesaurus (after deletion) is 19%, the lowest is 4%.

Table 1 summarizes the analysis results from the three different coding approaches: two different window sizes (3 and 19) and hand coding. The ultimate goal is to extract the key concepts that are strongly linked in each entrepreneur's self-presentation, and study the patterns of those connections across entrepreneurs. The values in Table 1 are the average unique and total concepts per text, and the average unique and total statements for the three approaches. Averages of unique concepts or statements are those that appear only once in a text; averages of total concepts or statements includes those that appear more than once in a given text. This is why the concept and statement unique averages are always smaller than the total averages.

Table 1. Concepts and statements by data reduction approach

	Concepts		Statements	
	unique	total	unique	total
Window size 3	11.6	20.2	29.4	37.3
Window size 19	11.6	20.2	24.3	24.8
Hand coding	17.6	25.7	15.2	15.5

The hand coding results identify more unique concepts (17.6) than AutoMap© (11.6), and more total concepts, although the total concepts AutoMap© finds are almost double the unique concepts (20.2). We speculate that this is because a human coder has a natural tendency to reduce redundancy, especially in categories where AutoMap© is likely to capture this redundancy (e.g. lists of educational credentials). A human coder also is more likely to first identify potentially meaningful words and phrases in the text, and then locate an appropriate concept in the thesaurus, instead of using the thesaurus to search for specific occurrences of words and phrases, therefore finding more unique concepts. In contrast, AutoMap© identifies many more total statements than a human coder (15.5), particularly at window size 3 (37.3). Again, AutoMap© seems to be capturing some redundancy, as there are 20% more total statements than unique statements (37.3 vs. 29.4). The high number of statements and redundancy at window size 3 is substantially reduced by using our modified operationalization of window size that creates statements based on where the concepts occur in the original text (24.8).

In the next stage of our analysis, we examine the statements formed by concepts within and across the five broad concept groups in the thesaurus. Understanding the following results requires more information about the content and structure of the thesaurus. The thesaurus has approximately 1150 entries which are translated into 76 concepts. The 76 concepts fit into five broad concept groups: personal description (the highest number of concepts), other people, the environment or external circumstances (the least number of concepts), success within and outside the organization, and discussion of their product's features.

Table 2 presents results for each of the three data reduction approaches. The percentages are based on the number of times that concepts in a concept group form a statement with concepts from that concept group versus with concepts from other concept groups. The most frequently occurring concept group, across all coding approaches, is the personal characteristics group, the least common is the environment or circumstance group. Personal characteristics concepts also consistently are the most likely to occur together. Across the coding approaches

there is substantial agreement for the first three concept areas about how statements combine concepts from different groups: personal characteristics, other people, and the environment or circumstances. There are larger differences between the approaches for the other two concept groups. Even the window size 3 and window size 19 results differ for concepts about success within and outside the organization. The approaches also generate quite inconsistent results for the fifth concept area (product discussion), most likely because of imprecise entries in the thesaurus and inconsistent coding of the thesaurus concepts by the coder.

Table 2. Percentage of statements formed by concepts from within vs. across different concept groups by data reduction approach

	Window size 3 results		Window size 19 results		Hand coding results	
Concept groups (CGs)	within CGs	across CGs	within CGs	across CGs	within CGs	across CGs
1. Personal characteristics	58%	42%	58%	42%	58%	42%
2. Other people	31%	69%	29%	71%	23%	77%
3. Environment; circumstance	9%	91%	7%	93%	11%	89%
4. Success in & out of the org.	18%	82%	28%	72%	38%	62%
5. Product discussion	18%	82%	21%	79%	44%	56%

Differences between the AutoMap© results and the hand coding results also occur in the most common statements found by each approach. The window size 3 and window size 19 results substantially agree about the most frequent statements (adjacent columns in Table 3). While the most frequent statements found by the hand coding have no overlap with those found by AutoMap,© this difference is not as dramatic as it first appears. First, only *eight* unique concepts form all of the statements in the AutoMap© coding results, compared to 14 unique concepts forming statements in the hand coding results, even though the total statements in the hand coding results are somewhat lower. The hand coding results and the AutoMap© results do show some consistency at the level of the individual concept, where five of the eight concepts in the AutoMap© results also occur in the hand coding results. Second, a closer examination of the AutoMap© statements indicate that it has (accurately) captured and coded a great deal of redundancy in these texts. For example, there are four reciprocal statements (e.g. education, education), indicating that two closely related words in a text were each translated into the same concept, and therefore into a statement. This type of redundancy is typically screened out by the human coder. Three pairs of statements are also complements (e.g. education, experience and experience, education) and can be considered essentially identical. The AutoMap© results thus generate a consistent, dense cluster of co-occurring concepts.

Table 3. Most frequently occurring statements by data reduction approach

Statements (concept, concept)	WS3	WS19
Education, education	27	25
Education, experience	11	7
Experience, education	7	4
Education, others	7	7

Statements	Hand coding
Education, training	8
Leadership, growth	8
Experience, expertise	7
Left, *entrepreneur*	7

Others, *education*	6	4
Entrepreneur, education	9	5
Entrepreneur, entrepreneur	7	7
Experience, *entrepreneur*	7	4
Entrepreneur, experience	7	5
Volunteer, volunteer	6	6
Others, others	5	6
Customer, employee	5	5
Entrepreneur, others	6	
Education, expertise		5

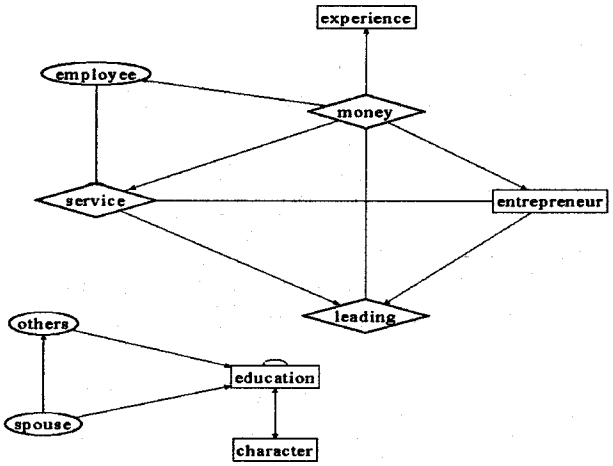
Money, growth	7
Rapid, growth	7
Community, *volunteer*	6
Volunteer, community	6
Entrepreneur, small	6
Experience, education	5
Growth, *entrepreneur*	5
Money, risk	5

Italics = concepts found by all coding approaches

The discussion so far has described how different coding approaches can alter the results of map analysis, even when a consistent coding scheme and a detailed thesaurus are used. We will now move on to the second question: what do these results tell us about the strategies that entrepreneurs use to manage their self-presentation? The map analysis results allow us to extract four specific self-presentation strategies: the prepared entrepreneur, the driven entrepreneur, the creative niche entrepreneur, and the humble entrepreneur. An entrepreneur may use several of these strategies within his or her profile. Each of the strategies serves a rhetorical purpose within the text, allowing the entrepreneur to position him or herself as an entrepreneur worthy of winning the award. The text of an entrepreneur's self-profile can be conceptualized as a series of distinct discursive moves, each of which is a link in the entrepreneur's map.

Figure 2 uses the set of concepts (10) and statements (17) in one entrepreneur's map to form a visual representation of a self-profile (window size 19). There are four personal characteristic concepts from cluster 1 (experience, entrepreneur, education, and character), but they are not the most central concepts, which are money and service, each linked to five concepts. This entrepreneur presents his personal characteristics as contributing to his success, but they are not as important in his self-presentation as the role he gives to good service and sound finances. This can also be seen in the disconnected subgraph, where the entrepreneur discusses people outside the company and his personal history.

Figure 2. Visualization of an entrepreneur's concept network (window size 19)



Concept Clusters: = cluster 1; = cluster 2; = clusters 3, 4, 5

The first strategy—*the prepared entrepreneur*—is the most obvious one, and can be seen in the dense concentration of concepts related to education and other general background qualifications. This is the most common strategy, and could almost be considered a "kernel" that all the other discursive moves in a profile build on as a base. All three coding techniques capture this strategy, and in Table 3 these concepts are linked to other concepts that characterize alternative strategies. The following excerpt is a typical example of an entrepreneur presenting himself as the prepared entrepreneur:

> "Undergraduate and graduate education in the Electrical Engineering disciplines have provided a thorough training in analytical thinking and problem solving. This was followed by formal MBA studies ..."

The entrepreneur mentions the specific areas of his educational background, and uses this to position himself as qualified to be an entrepreneur in his particular product area. In addition to emphasizing the length of his education by separating his schooling into two separate sentences ("undergraduate and graduate" and "formal MBA studies"), he uses the words "disciplines" and "thorough training" to underscore that this particular field is important for what he went on to do as an entrepreneur. This prepared entrepreneur has now placed himself in a position to act entrepreneurially through his education and training.

In contrast, *the driven entrepreneur* strategy focuses on personality attributes that link the entrepreneur to the characteristics typically associated with entrepreneurs, not to formal qualifications. This strategy is best captured in the AutoMap© coding by the concept "entrepreneur," and in the hand coding by the concepts linked to "entrepreneur," such as "left" and "small" (Table 3). A typical excerpt from an entrepreneur employing this strategy follows:

> "[Jane Doe's] entrepreneurial career has been marked by risk taking, perseverance, and [the] ability to dive in and grasp new business disciplines. She started [Company X] in 19xx. While [Jane Doe's] background was engineering, she hired an engineer and took charge of the marketing, sales, finance, purchasing and manufacturing areas."

The entrepreneur in this excerpt has pushed herself beyond the comfort boundaries that her background and expertise have prepared her for. The three personal characteristics attributed to her in the first sentence all emphasize this flexibility and adaptability required to survive as a small business with limited employees and resources. The explicit contrast in the third sentence emphasizes this point: instead of simply being a specialist, she is prepared to learn other business areas through "perseverance" not formal training. The second sentence also states her claim to being an entrepreneur—she started a company, the quintessential entrepreneurial move.

Authors using *the creative niche entrepreneur* strategy position themselves as a good entrepreneur because of how they positioned the company. The entrepreneur's claim to success is not based on either formal training or personality attributes. Instead, entrepreneurs using this strategy present themselves as having found success by coming up with something completely new, or by creating a product to meet a specific market need which they had the insight to see. In the following excerpt, an entrepreneur begins this third self-presentation strategy:

> "[John Doe] has a strong background in both technical and business issues, accompanied by a level of creativity that has helped to chart a rather unconventional -- yet highly successful -- course for [Company X]."

The contrast in this sentence is implicitly between the "strong background" (only a part of how the entrepreneur succeeded) and his "creativity" and "unconventional" actions as an entrepreneur. Balancing the latter two adjectives, the word "chart" emphasizes the role that the

entrepreneur's conscious positioning has played, implying direction and forethought. The qualifying clause at the end of the sentence is inserted between dashes to explicitly indicate that this strategy has been successful for the company. This may indicate the author believes the reader is less likely to accept this claim than to accept a claim based on the first two strategies.

The final self-presentation strategy, *the humble entrepreneur*, is a clear contrast with the rhetorical positions staked in all three of the previous strategies. Entrepreneurial success attributable to the entrepreneur's actions (their preparation, personality, and insight) are played down, and the role of others in the entrepreneur's life is played up. There are two specific substrategies: the first links the entrepreneur to others who may have played a role in the company's success, the second links them to others less fortunate than themselves. In Table 3, concepts representing both these substrategies are captured in all three coding approaches (e.g. community and volunteer). The following excerpt primarily employs the first substrategy:

> "The positive relationships [John Doe] has cultivated with employees and the community have been fundamental to [Company X's] success. [John Doe's] emphasis on employee participation has created a work environment in which all employees are members of the extended [John Doe] family."

In these two sentences, the entrepreneur positions himself as someone who is not successful alone, but because of how he treats others around him. In the second sentence, the author actually metaphorically encompasses the company's employees into the family of the entrepreneur's himself, extending the positive associations of a family to the business. Indeed, according to the author, this has been "fundamental" to entrepreneurial success, and something he has invested time in ("cultivated").

The results presented above lead to two broad conclusions about how to interpret the concept networks that are the results of map analysis. First, it is important to remember that the less data reduction, the more the original meanings built into the structure of the texts' sentences are preserved. Using only the concepts in the thesaurus yields a very stylized portrait of the entrepreneur's concept network. The different approaches to data reduction yielded slightly different conclusions about issues such as the density of networks and the frequency of specific concepts, as well as how the concepts in those areas related to concepts in other areas. However, the three approaches evaluated here were successful in allowing us to identify different self-presentation strategies. A specific contribution of this paper is our modified definition of window size 19 that allowed us to see how the concepts relate to each other within the structure of the entire text, rather than at the sentence or within sentence level. This redefinition allowed us to eliminate considerable redundancy otherwise captured in the automated coding.

While there are non-trivial differences between the coding results generated by different coding approaches, these differences highlight the substantial strengths of using either (or both) of the approaches. The advantages of automated coding are clear: it is fast, consistent, allows the researcher to make substantial data-reduction, and captures much of the same information as hand coding. The limitations of automated coding are the time required to build a thesaurus, and the inevitable possibility that it will miss meaningful information or capture information erroneously (Type 1 and Type 2 errors). The advantages of hand coding complement these weaknesses; the human coder can minimize both types of errors, and screen out redundant information when forming links between concepts. The disadvantages are that hand coding is much more time consuming, and there are many opportunities for inconsistencies as the coder's own interpretations of specific concepts emerges through reading multiple texts.

One value of examining the results of different approaches is the increased confidence it gives us that we have successfully identified distinct self-presentation strategies that these authors use in their texts. In additional analysis, the four self-presentation strategies identified in this paper can be further investigated and refined. For example, looking within each profile at the distribution of the concepts by clusters would indicate how dominant a cluster is within a given profile. Some profiles may focus on elaborating in particular areas, while others may present a balanced portrait. It would also be useful to know how often an entrepreneur employs multiple strategies. Visualizing more concept networks such as in Figure 2 would help identify how connected or disconnected the concepts are in a particular profile. Seeing the networks helps understand which concepts are central in a self-presentation, and which are more peripheral. In future work, as we analyze the entire dataset, we can refine the thesaurus, examine in more detail how the results of different coding choices influence our interpretations, and characterize additional strategies used in the entrepreneurs' self-presentations.

Computational research on texts is a growing area of research, but often the results of this research focus on the quantity of texts analyzed, and not on the quality of the interpretation of those analyses. The map analysis results and interpretation presented here demonstrate that understanding concept networks depends on the researchers approach to data reduction, and how he or she measures and defines a tie (to some extent the level of detail is dictated by the researcher's question). More specifically, the results here show that the automated analysis of texts can provide information broadly consistent with the results of human coding, and that this information about the concept networks in the texts allows us to make meaningful conclusions self-presentation strategies and their underlying patterns.

Bibliography

Carley, K. (1993). "Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis." *Sociological Methodology* 23, 75-126.

Carley, K. (1997a). "Extracting Team Mental Models Through Textual Analysis." *Journal of Organizational Behavior* 18, 533-558.

Carley, K. (1997b). "Network Text Analysis: The Network Position of Concepts." (Ch. 4). In Text Analysis for the Social Sciences, Carl W. Roberts, ed. Mahwah, NJ: Lawrence Erlbaum Associates, Inc. (79-102)

Corman, S., T. Kuhn, R. McPhee, and K. Dooley. (2001) "Studying Complex Discursive Systems: Centering Resonance Analysis of Organizational Communication." Working paper.

Danowski, J. (1993). "Network Analysis of Message Content." In W. Richards & G. Barnett (eds), Progress in Communication Sciences 12, 197-221. Norwood NJ: Ablex.

Roberts, C. ed. (1997). Text Analysis for the Social Sciences. NJ: Erlbaum and Associates.

Shapiro, Gilbert and John Markoff. 1997. "A Matter of Definition" (Ch. 1). In Text Analysis for the Social Sciences, Carl W. Roberts, ed. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.