Communication & Science

JOURNAL
2014 MAY 03

Woelfel, J. (2014). Procedures for Precise Text Analysis: Alternative Methods for Cluster Analysis. *Communication & Science Journal, 2014MAY03.* (received 2014APR24)

A RAH Press Publication
http://www.comSciJ.com

Procedures for Precise Text Analysis:
Alternative Methods for Cluster Analysis

Joseph Woelfel
University at Buffalo, State University of New York
April, 2014

jwoelfel@galileoco.com

"The object of all science, whether natural science or psychology, is to co-ordinate our experiences and to bring them into a logical system" – Albert Einstein

Introduction

The idea that the complexity of human experience is underlain by a small set of thema or archetypes stretches to antiquity, as is clearly visible in Plato's Ideas and Aristotle's Categories[1]. Early social scientists like Charles Spearman and L. L. Thurstone hoped to be able to reduce test scores of individuals to a small set of underlying "factors" defining human intelligence. The growth of "big data" has intensified efforts to find convenient methods of reducing massive amounts of data to smaller, simple, meaningful and useful representations.

One of the most popular methods has been categorization or clustering, in which various elements in the data are considered similar enough to be treated alike. Among the earliest such schemes is VALS, (values, attitudes and lifestyles), launched by SRI in 1978[2]. Subsequently many more VALS-like schemes followed. Although substantial differences exist, most follow a simple two-stage model, where individual's values, attitudes and lifestyles are measured in some way, similarities among them are calculated by some algorithm, and then these measurements are reduced to a smaller number of categories through some clustering or factor-analytic scheme. All the individuals in each resulting category are considered similar enough to be dealt with identically by marketers, advertisers, and other interests.

Although Spearman's and Thurstone's methods were based on individually administered tests, and early VALS-like systems were usually based on questionnaire-type data, by far the largest body of data available today consists of text, and the focus of contemporary "big data" research has shifted to the analysis of text, and this article will focus on text as well. Many programs and algorithms for text analysis exist, but it is convenient to divide them into two broad types: rule-based analysis and propinquity based analysis. Rule based text analysis consists of those type of analysis based on linguistic, syntactic, grammatical or other theoretical schemes of analysis. These methods differ among themselves widely, but share the notion that the correct interpretation of language depends on some rules or schemes, either learned or genetic[3, 4]. We will not be concerned with rule based clustering methods in this paper.

Propinquity based analysis, on the other hand, considers grammar, syntax, rules and the like to be devices invented by analysts rather than the basis on which individuals generate and interpret language. Instead, it assumes that words tend to become associated in meaning simply because they frequently occur "close" to each other in discourse. In this view, when President Nixon says, "…the American People deserve to know whether their president is a crook. Well, I am not a crook…" the "not" is meaningless, and the concepts of "Nixon" and "crook" are forever associated thereafter.

Propinquity analysis consists of two steps: first, the measurement of the propinquities or distances among the elements, and second, the procedures by which the elements are divided into categories or clusters once the propinquities have been established. We will consider here two broad types of clustering algorithms: hierarchical or "hard" clustering methods, which assign each element into its one "best" category, and

context-sensitive algorithms, which can assign the same element into one of several possible categories depending on context.

Measuring Propinquity

The oldest form of propinquity analysis is co-occurrence analysis. The most elementary form of this type of analysis is what Danowski[5] calls the "bag of words" approach[1]. In this form of analysis, the number of times words co-occur in the same "bag" – e.g., document, page, episode, utterance, etc. – is counted, and a matrix of frequencies of co-occurrence is computed. This co-occurrence matrix is the basis of all further analysis. Early examples of such programs are Danowski's Wordij,[8] Newton, which constructed co-occurrence matrices based on the co-occurrence of 150 behaviors in 15 second intervals of prime time TV shows in five countries;[9] and Catpac,[7] which has been used very extensively in a large number of substantive areas worldwide.

Although each of the various types of text analysis has its supporters, direct comparisons of different analysis routines on the same data are rare. In this paper, two of the most widely used similarity models – a co-occurrence model and a neural network model – are directly compared and contrasted on the same text. Further, certain important difficulties associated with the mathematical analysis of both methods are discussed, and an alternative model is presented.

Some examples:

Consider the following simple text:

Blue

I only had three dreams in my life I remember, and those none too clearly. I dreamed I went hunting with my father, and it was the only dream I ever dreamed in color. I don't remember anything about it, except the colors of the leaves and trees where we were hunting, and, if you made me swear to it, I couldn't in all honesty say I was sure we were hunting, or that my father was with me. I just remember the colors in the woods.
-1
I actually dreamed another dream in color, but only one color – blue. It was in the yard of our house, and it switched to my uncle's cottage in Sunset Bay. There were flying saucers flying around the yard/bay, and they were blue, and they were very compelling.
-1

[1] Woelfel refers to the bags as "cases" or "episodes", and the Catpac software and manual refer to the "bag of words" method as "case delimited mode."6. J. Woelfel, *Artificial neural networks for cluster analysis*. (RAH Press, Amherst, NY, 2009), 7. J. Woelfel, Journal of Communication **43** (1), 63-80 (1993).

I remember another dream about an attic in a house, which I now think might be my mother's parents' house, and I confuse that with another dream about a huge house with many hidden and mysterious rooms and stairways, but it's terribly vague, and not in color. The emotion is there, though, and it's an emotion of strangeness and discovery, a tingling emotion, as are the other two dreams I remember.
-1
Walter Mosley's Blue Light made me think of these dreams. He's a far better writer than I am, and his book starts with rays of blue lights on a much more cosmic scale than my dreams – rays of blue light from Neptune crashing through the sun to earth – awakening questions in frogs and murder in young men. I still have no idea what his blue lights might be, but I know that I know the blue lights from my childhood dreams. I know the blue lights, and I can still feel them.
-1
When I was a teenager I made a device that emitted blue light. It was like the cardboard tube from a roll of paper towels, with a blue light bulb, like an old fashioned Christmas tree light – a Mazda – in it. I connected it to a cardboard box with a few resistors and capacitors wired in no particular way. I knew it had no effect on anything, but I called it a "synapsifier", a fictitious device that worked on the synapses of the human brain. I check the frontispiece of Walter Mosley's book, and find the copyright is 1998. My synapsifier filled my attic room with blue light over fifty years earlier – about the time of Blackboard Jungle, plus or minus…

So I had the blue light thing early, but I didn't use it.
-1

This text consists of five paragraphs delineated by the marker "-1". Each marker delineates one "bag of words" or "episode"." A simple cluster analysis utilizing the co-occurrence method[2] yields the dendogram shown in Figure 1:

[2] All analysis in the following examples were performed using CatpacIII™ from The Galileo Company.

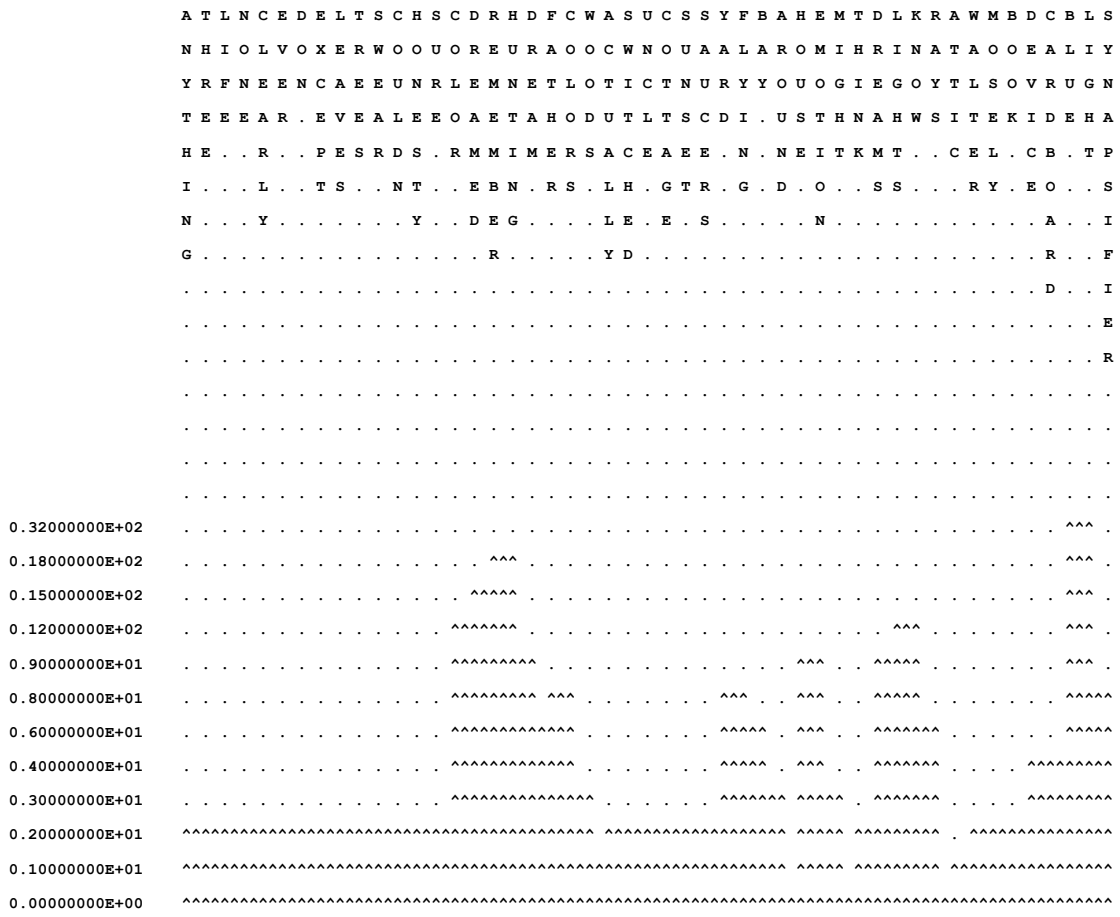

Figure 1: simple co-occurrence analysis

Figure 1 shows a fairly shallow dendogram with only eight levels (as shown by the "pseudo z-scores" at the left side of the Figure produced by the Johnson's Hierarchical Clustering algorithm in CatpacIII). Nevertheless, there is enough signal in the result to show clearly the father/hunting episode, the color flying saucer dream (although the saucers are omitted), the Walter Moseley book and the "synapsifier device, both of which are sub-clusters of the Mosley/synapsifier cluster. There is also a fifth cluster, the dreams/rays/lights cluster.

A significant improvement on the "bag of words" approach came from Danowski[10] who first introduced the "sliding window" approach. His approach introduces a better measure of propinquity, in that words that are close together will co-occur more often in the sliding window than words that are further apart. Specifically, in a two-word window, only contiguous words will co-occur. In a three-word window, contiguous words will co-occur twice, while words separated by one word will co-occur once. In a seven-word window, contiguous words will co-occur six times, words separated by one word will co-occur five times, and so on until words separated by five words will co-occur once.

Figure Two shows the results of a co-occurrence analysis of the same text, this time using a moving window of size three, (the size Danowski recommends as optimal). Figure two does not appear to be an improvement on the simple "bag of words" co-occurrence model for this particular text, yielding more and smaller two-word clusters, such as dreams/remember, color/emotion, dream/dreamed, hunting/father, Walter/Mosely, synapsifier/device and the like, and a few three-word cluster such as blue/light/cardboard, might know lights and the like. The number of levels is also reduced to six.

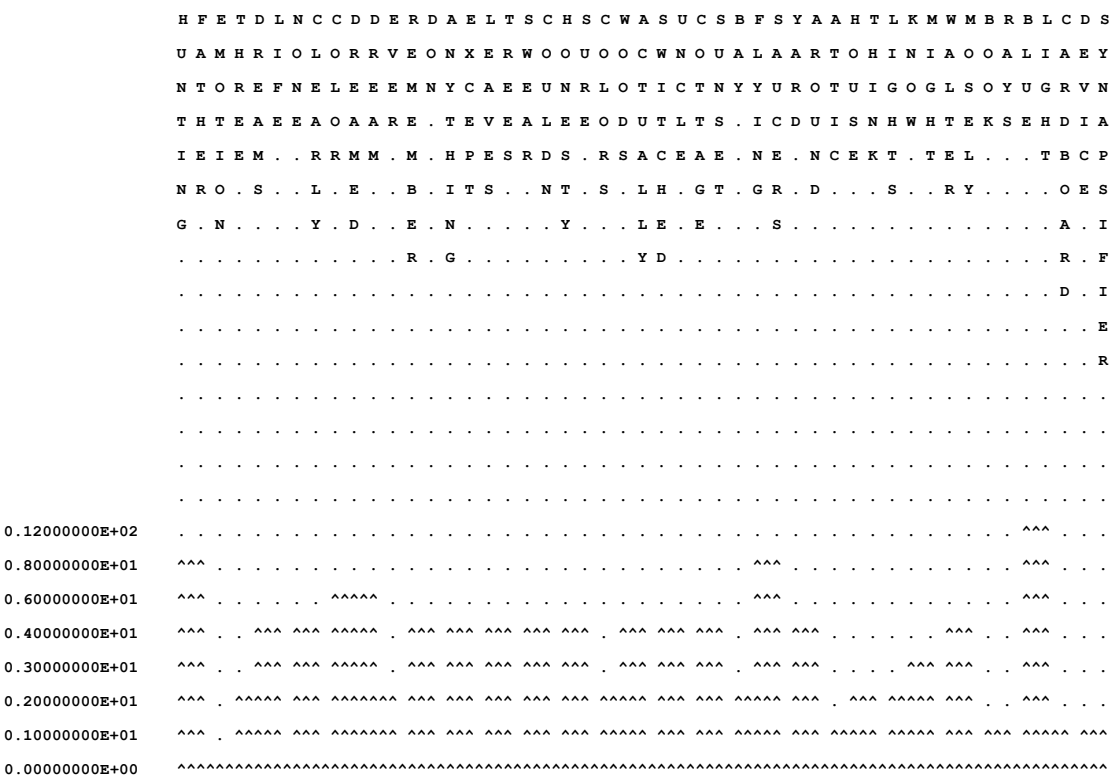

Figure 2: Co-occurrence model with 3 word sliding window

Increasing the window size to seven, does increase the depth and detail substantially, yielding more than twice as many (15) levels, four and five word clusters, and even sub-clusters within clusters.

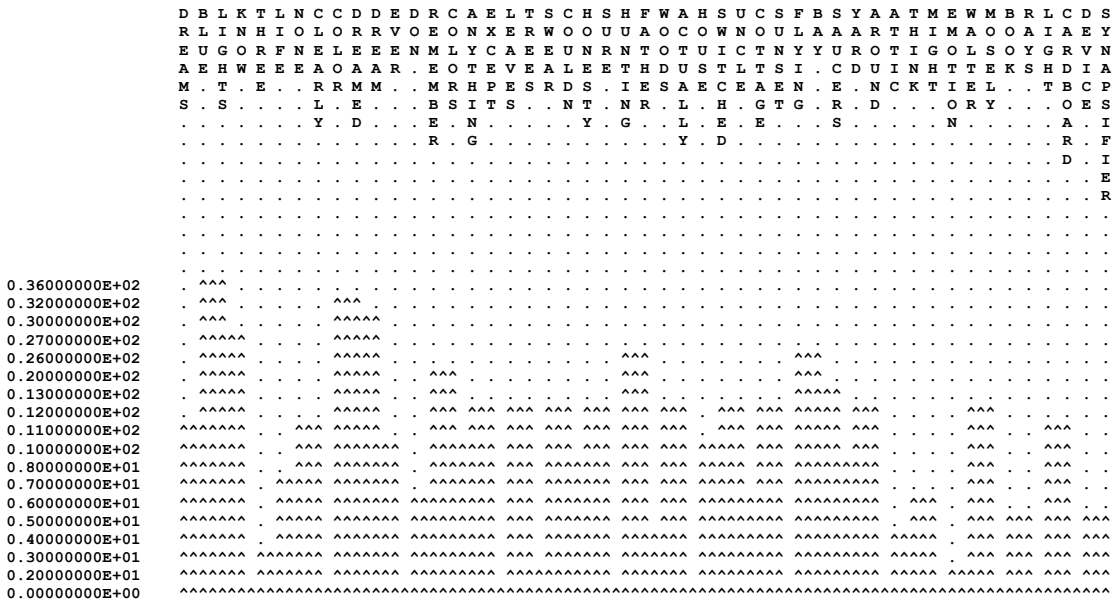

Figure 3: Co-occurrences with 7 word sliding window

More recently Woelfel[7] introduced an artificial neural network into text analysis. This network was implemented in Catpac™, which was originally a program which computed co-occurrences using the "bag of numbers" approach, where the beginning and end of each "bag" was determined by codes embedded in the text. Later, Catpac implemented Danowski's sliding window method, and, in the late 1980's, introduced a single pass unsupervised neural network to measure the propinquity relationships among the words in the text.

Initially, the network creates a set of artificial neurons, one for each word in the text. Then it "activates" those neurons whose associated word is in the sliding window at each iteration. The "connections" (degree of closeness or propinquity) of those neurons that are co-present in the window are then incremented. As the network grows, however, and connections are established among the neurons, activation of neurons in the window can result in the activation of other neurons not in the window, which are positively connected to those in the window. Connections among all these neurons are also incremented[3].

What this means in practice is that the propinquity of nodes is not established simply on the basis of pairwise co-occurrences, but on the basis of both direct pairwise relations and complete n-way indirect relationships among all the nodes. In practice, the result is deeper, more finely detailed relationships among the nodes, as shown by deeper, more detailed dendograms and perceptual maps.

[3] Complete operation of the network, including forgetting, normalization and other issues is not discussed here. For more detail, see Woelfel, 1993.)

Figure 4 shows the results of a neural network analysis of the same text, using a seven-word window:

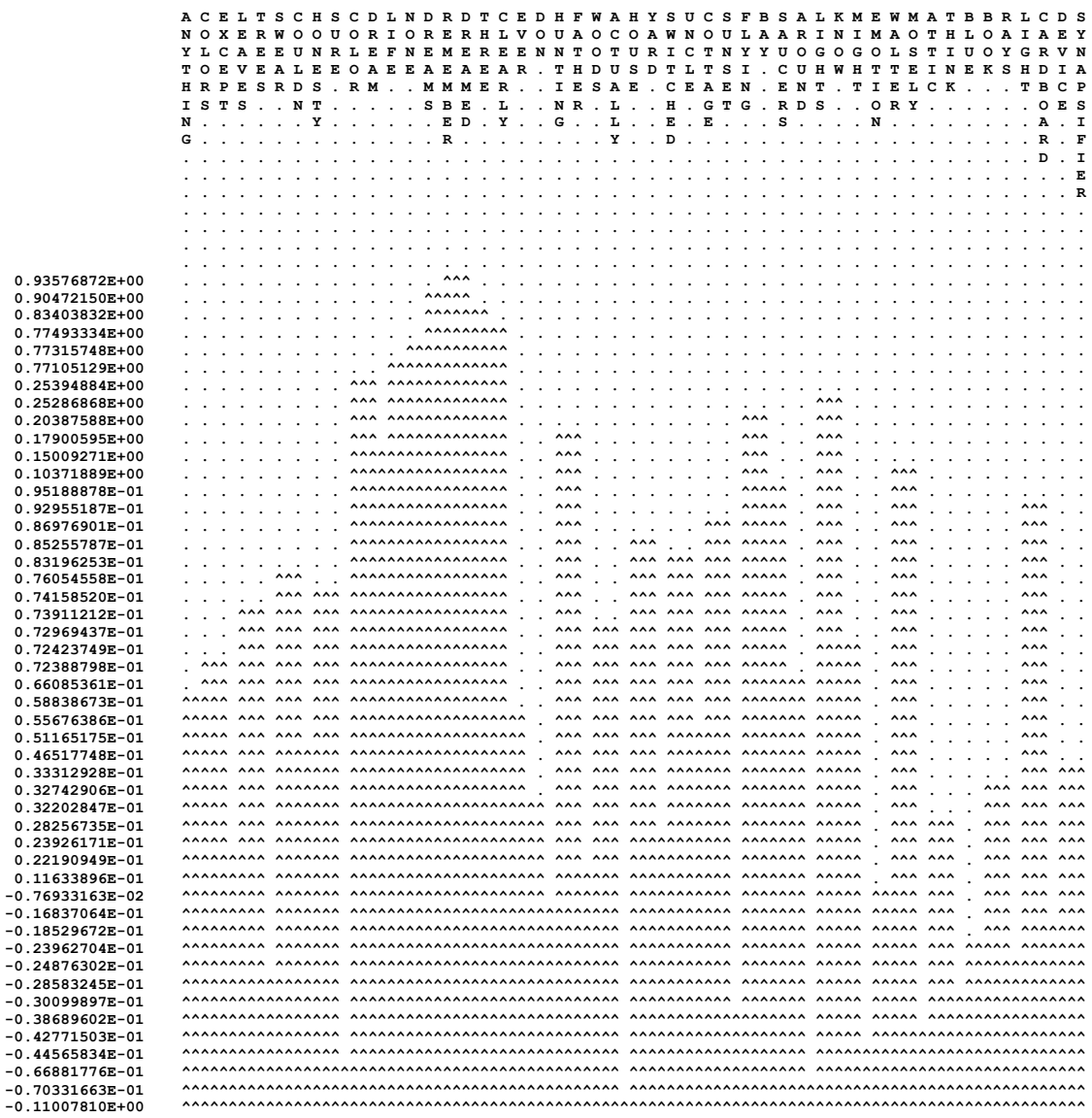

Figure 4: Neural network with 7 word window

Figure 4 shows that the neural network with a 7-word window clearly finds the deepest, most finely articulated structure of any of the methods tried so far. With 33 distinct levels of clustering and many deep clusters subdivided into smaller sub-clusters. What may be lost in the mix, however, are the four original episodes – the three dreams and the Moseley book and the synapsifier.

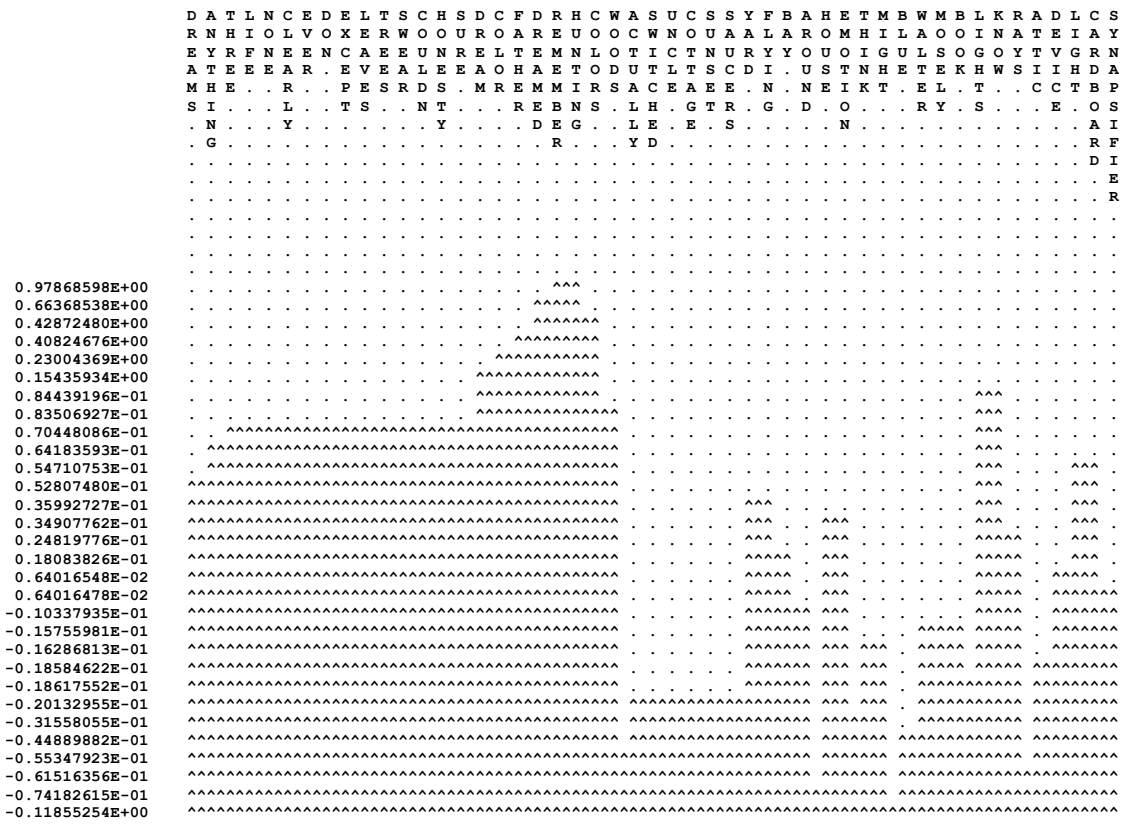

Figure 5: Neural Network with "Bag of Words" or Episodes

Figure 5 shows the neural network solution using the original episodes or bags. The three dreams and the Moseley/synapsifier episode are clearly visible in the solution, although the detailed picture of sub-clusters is lost. Generally, if the episodes or "bags" are not meaningful, the moving window neural network is the method of choice. But if the episodes or bags are substantively meaningful, the episodic of "bag of words" approach should be preferred. Since the actual run-time for the software is usually only a second or two, there is, of course, no reason both procedures shouldn't be tried and reported.

One of the instances in which the "bag of words" or episodic method appears more appropriate is shown in the case of medical diagnosis and treatment. Each disease may well be considered a category or cluster that contains symptoms, diagnostic tools, diagnoses, treatments and outcomes. Figure 6 shows the symptoms and suggested diagnostic tools associated with several diseases in the Merck Manual Online[11].

ABDOMINAL PAIN
WAVES OF DULL PAIN WITH VOMITING
INTESTINAL OBSTRUCTION
ULTRASOUND
-1
ABDOMINAL PAIN
COLICKY PAIN THAT BECOMES STEADY
APPENDICITIS
STRANGULATING INTESTINAL OBSTRUCTION
MESENTERIC ISCHEMIA
ULTRASOUND
-1
ABDOMINAL PAIN
RECURRENT
ULCER DISEASE
GALLSTONE COLIC
DIVERTICULITIS
MITTELSCHMERZ
-1
ABDOMINAL PAIN
SHARP, CONSTANT PAIN, WORSENED BY MOVEMENT
PERITONITIS
-1
TEARING PAIN
DISSECTING ANEURYSM
ABDOMINAL PAIN
-1

Figure 6: Five abdominal diseases from the Merck Online Manual.

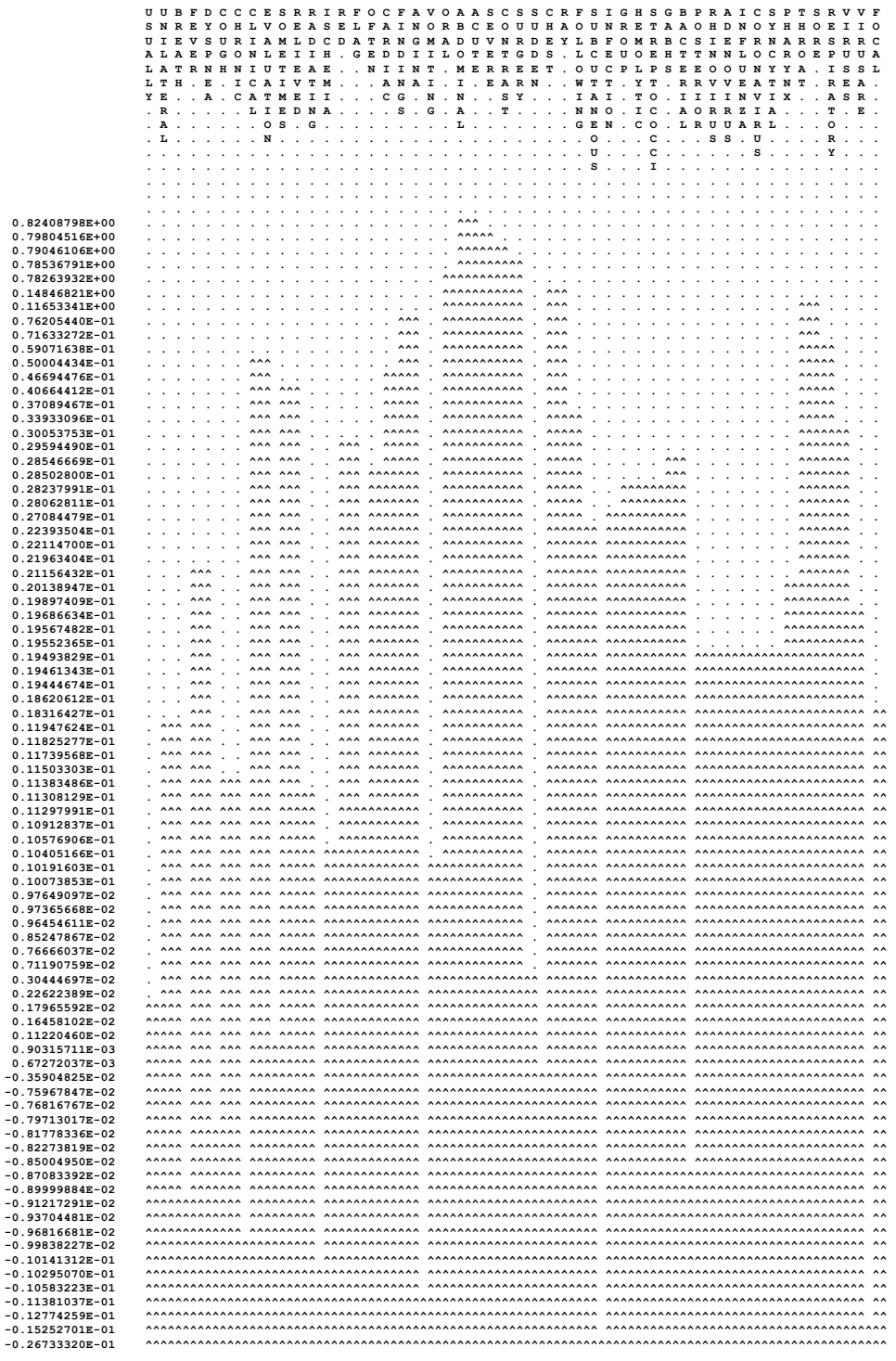

Figure 7: Neural Network with Episodic or "Bag of Words" Analysis (Merck Data - abridged[4])

[4] Due to space considerations, only a small subset of the Merck data was included in the analyses shown here.

Since the Merck data consists of a list of symptoms, diagnoses and proposed diagnostic procedures, it would seem that the episodic or "bag of words" approach would be appropriate, since a moving widow would necessarily lump symptoms, diagnoses and treatment strategies of one disease with it's neighbor in the data file every time the moving window slid across the end of one disease and the beginning of another.

But Figures 6 and 7 reveals a deeper problem: the hierarchical clustering methods discussed so far all share the same shortcoming: they require that every element be placed into one and only one "best" cluster. *Vomiting*, for example, is clustered with angina in Figure 7, but clearly vomiting is a symptom of many different diseases. Similarly, as Figure 6 shows, *abdominal pain* occurs in many diseases, and obviously can't be classified into only one best cluster.

Since the same symptoms can and do occur in several diseases, any clustering method that requires that each element be assigned to one and only one "best" cluster is inappropriate. Fortunately, the neural algorithm can discover content-sensitive clusters that allow elements to belong to more than one cluster at the same time, and to different clusters depending on context.

Because the neural network constructs a network of neurons through not only direct links, but through all n-way indirect links as well, it can calculate the degree to which any given neuron will be activated when any other subset of neurons is activated. This model is implemented in Indstar™, a neural network for non-hierarchical cluster analysis[12]. Basically, Indstar can answer the question "what other elements will reside in a category that includes any given set of elements?" So we can query Indstar by activating the neuron corresponding to *abdominal pain*, and it will respond by activating all the other elements that are (sufficiently) linked to it:

```
Enter as many items as you want, Doctor. -1 when done

ABDOMINAL PAIN

          And the winners are...            Activation


                ABDOMINAL PAIN                  1.0000
                ACUTE                           0.0052
                SEVERE                          0.0052
                ELDERLY                         0.0026
                CT WITH ORAL CONTRAST           0.0102
                SURGERY                         0.0052
                INFANT                          0.0026
                MILD                            0.0026
                INCONSEQUENTIAL                 0.0026
                HIV                             0.0026
                IMMUNOSUPPRESSANTS              0.0026
                CORTICOSTEROIDS                 0.0026
```

FURTHER TESTING	0.0026
VISCERAL PAIN	0.0051
VAGUE	0.0051
DULL	0.0051
NAUSEATING	0.0051
POORLY LOCALIZED	0.0051
UPPER ABDOMEN	0.0026
STOMACH	0.0026
DUODENUM	0.0026
LIVER	0.0026
PANCREAS	0.0026
FOREGUT	0.0026
LOWER ABDOMEN	0.0025
DISTAL COLON	0.0025
GU TRACT	0.0025
HINDGUT	0.0025
ACUTE WAVES OF SHARP CONSTRICT	0.0026
RENAL COLIC	0.0026
BILIARY COLIC	0.0026
ULTRASOUND	0.0076
WAVES OF DULL PAIN WITH VOMITI	0.0025
INTESTINAL OBSTRUCTION	0.0025
COLICKY PAIN THAT BECOMES STEA	0.0025
APPENDICITIS	0.0025
STRANGULATING INTESTINAL OBSTR	0.0025
MESENTERIC ISCHEMIA	0.0025
RECURRENT	0.0025
ULCER DISEASE	0.0025
GALLSTONE COLIC	0.0025
DIVERTICULITIS	0.0025
MITTELSCHMERZ	0.0025
SHARP, CONSTANT PAIN, WORSENED	0.0025
PERITONITIS	0.0025
TEARING PAIN	0.0025
DISSECTING ANEURYSM	0.0025
SUDDEN ONSET	0.0025
PERFORATED ULCER	0.0025
RENAL STONE	0.0025
RUPTURED ECTOPIC PREGNANCY	0.0025
TORSION OF OVARY	0.0025
TORSION OF TESTIS	0.0025
SOME RUPTURED ANEURYSMS	0.0025
FLAT AND UPRIGHT ABDOMINAL X-R	0.0025
UPRIGHT CHEST X-RAYS	0.0025

Figure 8: Indstar™ analysis of nodes connected to *abdominal pain*.

Clearly, as Figure 8 shows, *abdominal pain* is a symptom of alternative possible diseases depending on, among other things, whether it is acute, severe, visceral, vague,

dull, nauseating, and/or poorly localized, and classifying it as a member of its one "best" disease is a serious error. Which disease that symptom is indicating in any given case is determined by the context, which will consist of other symptoms and test results presented. Figure 9 shows how Indstar determines to which category *abdominal pain* should be assigned in the context of an additional symptom, *colicky pain that becomes steady*:

```
Enter as many items as you want, Doctor. -1 when done

ABDOMINAL PAIN

COLICKY PAIN THAT BECOMES STEADY

          And the winners are...            Activation[5]

                ABDOMINAL PAIN                  1.0000
                ACUTE                           0.0052
                SEVERE                          0.0052
                CT WITH ORAL CONTRAST           0.0102
                SURGERY                         0.0052
                VISCERAL PAIN                   0.0051
                VAGUE                           0.0051
                DULL                            0.0051
                NAUSEATING                      0.0051
                POORLY LOCALIZED                0.0051
                ULTRASOUND                      0.0101
                COLICKY PAIN THAT BECOMES STEA 1.0000
                APPENDICITIS                    0.0050
                STRANGULATING INTESTINAL OBSTR 0.0050
                 MESENTERIC ISCHEMIA             0.0050
```

Figure 9: Indstar™ neural activations for *abdominal pain* and *colicky pain that becomes steady*.

Figure 9 shows that including the additional symptom *colicky pain that becomes steady* into the context of *abdominal pain* changes the categories to which it "belongs". Several possible diagnoses have been activated by the inclusion, such as appendicitis, strangulated intestinal obstruction, mesenteric ischemia and the like. Several suggested diagnostic tools have also been activated, such as ultrasound and CT with oral contrast. Further, many possible diagnoses and symptoms have been *deactivated* by the additional context. Which diagnostic tools to use and which disease ultimately turns out to be correct will depend on still more contextual cues. What is important, however, is

[5] For reasons of space, elements with activation values below .005 are omitted.

that *the category to which any given element "belongs" is dependent on context, and cannot be settled once and for all by a simple hierarchical clustering scheme*.

Conclusions

Hopes to reduce the complexity of human experience to a small set of meaningful categories date to antiquity. The earliest thinkers attempted to discover these underlying archetypes by reasoning, and developed many theoretical schemes – a practice which continues unabated today. Beginning in the early 20th Century, mathematical and later computer-based techniques devolved. Most of these involve the measurement and calculation of similarities among elements by some scheme, followed by a mathematical reduction of the rank of the similarities to some underlying set of factors, clusters or other archetypes.

A wide variety of such computational/mathematical techniques continue to exist side by side, but this paper has shown that the method by which the similarities among elements is measured and/or calculated is of fundamental importance, and so not all existing techniques ought to be considered equivalent. Methods appropriate for some types of data maybe unsuitable for others and vice versus.

Of even greater importance, the idea that a clustering scheme can be devised in which each element can be assigned to its one and only "best" category is shown to be frequently impossible due to the effects of context. A context-sensitive neural network model, Indstar, is presented which makes it possible to classify elements based not on some "inherent" meaning, but on the basis of the context in which they are experienced.

References:

1. P. Studtmann, in *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta (Summer, 2013).
2. D. Elgin and A. Mitchell, Co-Evolution Quarterly (Summer) (1977).
3. N. Chomsky, *Syntactic Structures*. (Mouton, The Hague, 1957).
4. J. R. Searle, The New Yor Review of Books **18** (112) (1972).
5. J. A. Danowski, in *Progress in Communication Sciences XII*, edited by G. A. Barnett and W. Richards (Ablex, NJ, 1993), pp. 197-222.
6. J. Woelfel, *Artificial neural networks for cluster analysis*. (RAH Press, Amherst, NY, 2009).
7. J. Woelfel, Journal of Communication **43** (1), 63-80 (1993).
8. J. A. Danowski, in *Communication Yearbook 5*, edited by M. Burgoon (Transaction Books, New Nrunswick, N.J., 1982).
9. B. Newton, E. Buck and J. Woelfel, Human Organization **45** (2), 162-170 (1986).
10. J. A. Danowski, TREC, 5 (1992).
11. R. S. Porter, MD and J. L. Kaplan, MD (Eds), (Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc, Whitehouse Station, N.J., U.S.A., 2004-2012).
12. B. Battleson, H. Chen, C. Evans and J. Woelfel in *Sunbelt XXVII, INSNA Social Networking Conference* (St. Pete Beach, FL, 2008).