A Nonhierarchical Neural Network Approach For Analyzing Textual Data

Brenda L. Battleson, Hao Chen, Carolyn Evans, Joseph Woelfel
Department of Communication, University at Buffalo (SUNY), Buffalo, New York 14260 USA

Introduction

Artificial neural networks excel at recognizing patterns in textual data. Its pattern recognition capability allows a neural network engine to assign weights representing the multiple connections among concepts. These weights can then be used to create *dendograms* or otherwise categorize concepts in a hierarchical manner. However this approach has its limitations.

Words often have different meanings depending on the context in which they occur. Yet a hierarchical clustering method is unable to fully describe multiple relationships because it is only able to show concepts connected in one way. Each concept is assigned to only one "best" cluster in the output suggesting that there is only one meaning of that concept in the data analyzed.

The use of a nonhierarchical approach can address this limitation since it allows the researcher to interact with the neural network to explore all possible meanings of a concept. Thus, in the resulting output a concept may appear in as many clusters as are appropriate.

In this study a large dataset containing multiple newspaper articles, is examined. Hierarchical and nonhierarchical procedures are compared.

Method

Opinions about the terrorist attacks of September 11, 2001 were of particular interest during the five-year anniversary of the event in September 2006. To gauge opinion, editorials, opinion pieces and letters to the editors of all U.S. newspapers indexed in the FACTIVA™ database were retrieved for the month of September 2006.

The 3.2MB text file was analyzed using the CATPAC™ text analysis program. Output consisted of scalar products matrix used to generate an artificial neural network (ANN) with output consisting of a weighted input network (WIN) file and the hierarchical clusters represented in both dendogram and 3-dimensional coordinate files. The ORESME ™ software was then used for nonhierarchical analysis of the CATPAC results.

Unlike the traditional forward feed-back propagation neural networks, ORESME[1] is an interactive activation and competition network, and any neuron can be an input, hidden or output neuron. According to Woelfel, the most appropriate representation of ORESME's schematic is Patrick Lemmens' interpretation of a neural network (Fig.1A.)

Figure 1: "Neural network" by artist Patrick Lemmens

In this study, the researcher assigned an input (activation) value to one or more terms and the resulting clusters of "activated concepts" in the ANN were compared. The nodes of an ANN are connected to each other by weights which represent their relative "closeness" in the network. They communicate with each other by a simple linear threshold rule:

The signal sent from any node *i* to any node *j* equals the product of the activation value of *i* and strength of the connection between *i* and *j*. Thus the total signal received by any node *j* will be the sum of the signals received from all the other nodes, or

$$anet_i - \sum_{j=1}^{N} w_{ij} a_j$$

[1] CATPAC™ and ORESME™ are PC versions of KAMANDU™ and LISTIAC™ (respectively) both of which are running on mainframe computers at the University at Buffalo's Center for Computational Research.

Results

HIERARCHICAL CLUSTERING

The dendogram and 3-dimensional map resulting from the CATPAC analysis show definite clusters:

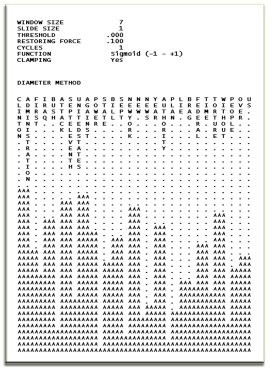

Figure 2. Dendogram of 2006 op-ed pieces in U.S newspapers for the month of September, 2006

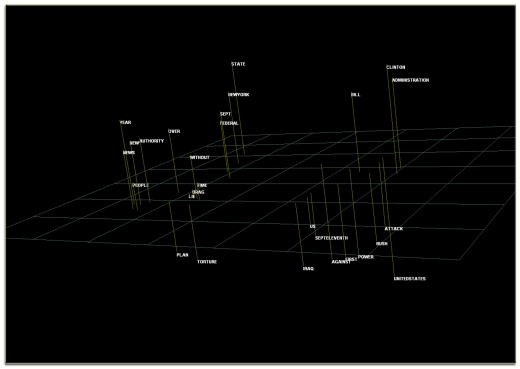

Figure 3. ThoughtView 3D map of hierarchical clusters identified using CATPAC.

NONHIERARCHICAL CLUSTERING:

ORESME allows the researcher to input or "activate" the neuron representing a concept of concepts. Multiple cycles allow for "learning" and weight adjustment of associated neurons. A threshold value determined by the researcher determines what associated neurons are activated. The results are new and often very different clusters of concepts.

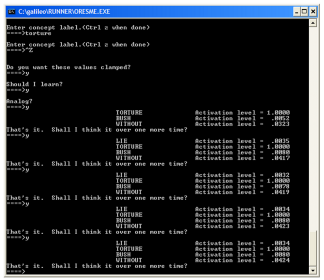

Figure 3. ORESME run showing results of analysis of the concept "TORTURE" through 5 cycles.

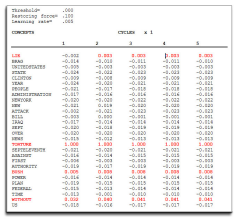

Figure 4. Output of ORESME run showing the activation levels of all concepts when an input activation level of 1.0 is assigned to the concept "TORTURE." Those concepts exceeding the threshold of .000 are activated. Terms clustering with TORTURE are LIE, BUSH and WITHOUT.

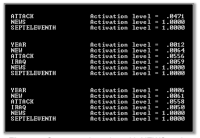

Figure 5. Concepts clustering with NEWS and SEPTELEVENTH.

The researcher is not limited to analyzing one concept at a time. Multiple concepts may be initially activated. (Fig. 5 & 6)

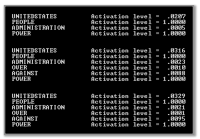

Figure 6. Terms clustering with PEOPLE and POWER.

Conclusions

The human brain is the most sophisticated example of a parallel distributed processing machine. The language used to express human ideas, attitudes and emotions is evidence of this sophistication. Yet we try to analyze language and human communication with bounded linear methods.

In this study, ideas generated with regard to a single commonly experienced event produced some predictable results. We see in the hierarchical cluster analysis that many concepts were grouped as anticipated—SEPTELEVENTH clustered with concepts like ATTACK, BUSH, UNITEDSTATES, IRAQ and most importantly US. Additionally, there was a TORTURE cluster, a NEWS cluster and even a BILL CLINTON cluster. However, this obviously represented only part of peoples' thinking with regard to 9/11 and this event's fifth anniversary.

The disadvantage of hierarchical cluster analysis is that we see only part of the picture. Concepts are placed in one "overall best fit" cluster when in reality, they can be in one or many clusters depending on such variables as context, time, place, etc. On the other hand, a nonhierarchical approach allows us to see some of those relationships that may not have been statistical "best fits," but are nonetheless important in finding meaning in the text. Consider the concept SEPTELEVENTH which has a clearly defined cluster illustrated in Fig. 3. When it is paired with another term, NEWS, the concepts with which it was originally clustered are not as important. Furthermore, terms with which it was not seemingly related are now much closer. Indeed, SEPTELEVENTH has multidimensional meaning.

This study is evidence that software like ORESME™ can be used to analyze text in a more meaningful way. There are quirks in the software and the output could be more user friendly. But this is of minor concern given the overall results of this research which support the need for a nonhierarchical approach. There are so many confounding variables when it comes to studying human communication. That it is impossible to control these variables only reinforces the importance of using nonhierarchical analysis to discover meaning.

References

Barnett, G.A. & Woelfel, J. (1979). On the dimensionality of psychological processes. *Quality and Quantity, 13*, 215-232.

Jain, A.K., Murty, M.N. & Flynn, P.J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3), 264-323.

Mangiameli, P., Chen, S.K. & West, D. (1996). A comparison of SOM neural network and hierarchical clustering methods. *European Journal of Operational Research, 93*, 402-417.

McClelland, J.L & Rumelhart, D.E. (1988*). Explorations in Parallel Distributed Processing : A Handbook of Models, Programs, and Exercises.* Cambridge: MIT Press.

Mead, G.H. (1934). *Mind, Self and Society From the Standpoint of a Social Behaviorist* (C. Morris, Ed.). Chicago: University of Chicago Press.

Ripley, B.D. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society B, 56*(3), 409-456.

Woelfel, J. (1995). Attitudes as nonhierarchical clusters in neural networks. In G. A. Barnett & F. J. Boster (Eds.) *Progress in Communication Sciences,* Vol. 13, Greenwich, CT: Ablex Pub. Corp., 213-227.

Woelfel, J. (1993). Galileo ORESME: User Manual. Royal Oak, MI: Terra Research and Computing.

Woelfel, J. & Stoyanoff, N.J. (1993). *CATPAC: A Neural Network for Qualitative Analysis of Text.* Paper presented at the Australian Marketing Association annual meeting, Melbourne, Australia.

More information

Please contact ***blk@buffalo.edu***. More information on this and related projects can be obtained at ***http://www.informatics.buffalo.edu/faculty/woelfel***. A PDF-version of the poster will soon be available in the literature section of the above site.

University at Buffalo
The State University of New York