Application of Text (Idea) Mining to Internet Surveys: Electronic Capture of the Structure of Ideas and Semantic Concepts

Jeffrey E. Danes
Orfalea College of Business
California Polytechnic State University
San Luis Obispo, California, 93405, USA
jdanes@calpoly.edu

Abstract

This paper demonstrates a quick and efficient method of assessing the ideation structure for a group of people via the Internet and text mining. Data collection using the Internet is increasing; although the Internet has made access easier figuring out what people think is still a challenge. Email was used to contact and then direct survey respondents to a web site. At the web site, open-ended questions requiring text (type written) responses were asked. Conceptual (ideation) structure was obtained via an algorithm similar to that suggested by Quillian [7] and Bonnet [1]. To discover ideation structure, a modified Hopfield neural network based text-mining algorithm was used to obtain the statistical weights of ideas and concepts and the weights of the joint occurrences of the ideas and concepts with other ideas and concepts. Applying neural network technologies to text allows the analysis of open-ended responses without incurring the expensive, time-consuming and error-prone task of manually reading the open-ended comments. The Internet via email contact and web-based survey input dramatically speeds the process.

1. Introduction

Data are typically collected through surveys conducted by phone or through the mail, consumer interviews in malls, and focus groups. The increasing use of the Internet for data collection has made access easier; however, figuring out what people think is still a challenge. This paper demonstrates a quick and efficient method of assessing the ideation structure for a group of people.

This paper uses a combination of electronic methods to measure the ideation (concept) structure of university faculty interested in developing communities. Email was used to contact and the direct faculty to a web site. At the web site, open-ended questions requiring text (type written) responses were asked. Conceptual structure was obtained via algorithms similar to those suggested by Quillian [7], Bonnet [1] and Chung, Pottenger, and Schatz [3]. The goal of the application of methods was to gather "collective wisdom" as quickly and efficiently as possible.

The content of the survey focused on thoughts, suggestions, and opinions regarding the startup and operation of a community development center at a major U.S. University. The survey focused on four textual questions; only one of these is reported in this paper: The faculty and staff's suggestions on how to encourage community development activities on campus and local community. This study used E-mail and a Web site to conduct a survey, designed to capture via open-ended (text) responses, the structure of ideas and semantic concepts.

Faculty (and administrative and support staffs) were contacted by email and directed to a web survey site. The survey contained four open-ended (text) questions including current projects and activities related to community development, projects planned for the next year, interests in community housing projects, and suggestions regarding the promotion of community development activities. Specifically, this last question asked for suggestions about how to encourage/promote community development activities on the university campus.

2. Electronic Data Collection

Four weeks before the end of the spring term 2002, an email message was sent to all faculty and administrative staff introducing the respondents to the survey topic and then directing them to a web page, the survey's URL. Two weeks later a reminder email was sent. One hundred and forty faculty and administrative staff responded to the survey request. The responding faculty and staff, of course, do not provide a representative sample of the entire faculty; however, they do represent those faculty and staff whom are highly interested in community development activities.

3. Discovery of Ideation Structure

Implementation of the analysis was accomplished via Text Analyst 2.0; the general idea of applying a modified Hopfield network to text is given in Chung, Pottenger, and Schatz [3].

The goal of the text analysis is to filter out meaningless elements, process the significant information, and identify the key semantic concepts contained within the text. Important concepts and word combinations from the text are identified by their frequencies of occurrence and joint occurrence with other concepts. The text is taken as a sequence of ideas organized by sentences into words. This sentence-string of words is moved through a window of variable length of text fragments. The goal is to discover ideation structure of the text. A modified Hopfield neural was used to obtain the statistical weights of ideas and concepts and the weights of the joint occurrences of the ideas and concepts with other ideas and concepts. The Hopfield network evolves by changing the weights assigned to the nodes and links between them into a stable configuration corresponding to the minimum of an energy function characterizing the semantic network. The ideation structure represents a conceptually accurate and concise map of the key ideas contained in the input text.

The modified Hopfield neural network is used to obtain a graph-like structure of statistical weights of concepts and the weights of the joint occurrences of these concepts with other concepts. The Hopfield network evolves by changing the weights assigned to the nodes and links between them into a stable configuration corresponding to the minimum of an energy function characterizing the semantic network. The semantic network represents a linguistically accurate and concise picture of the analyzed text. Applying neural network technologies to text allows the analysis of open-ended responses without incurring the expensive, time-consuming and error-prone task of manually reading the open-ended comments.

3.1 Modified Hopfield Algorithm

The Hopfield network is a neural network that may be viewed as a content-addressable memory structure. Knowledge and information is stored in single-layered interconnected neurons (or nodes) operating as memories representing patterns stored in the network. Algorithmic steps for the general application of a modified Hopfield algorithm to semantic networks are presented in Chung, Pottenger, and Schatz [3]. Their method is used below to illustrate the operations of Text Analysis 2.0.

3.1.1 Preprocessing

Preprocessing is a language dependent activity. It involves removal of all ancillary and commonplace words, words that carry no semantic meaning. It also involves the identification of stems of the words, while separating prefixes and suffixes. Additional prepossessing includes pooling words with common stems, "mean" and "meaningful" for example.

3.1.2 Assigning Synaptic Weights

Similarity of any two concepts is given by their co-frequencies of usage relative to individual frequencies, similar to an index of statistical covariance. The concepts generated by similarity analysis serve as the trained network. The concepts represent nodes in the network and the similarities, computed based on co-occurrence analysis, represent asymmetric semantic weights between ideas or concepts (nodes). The synaptic weight from node i to node j is denoted as w_{ij}. The relationship between each pair of ideas (nodes or neurons) is expressed as the real-valued asymmetric similarity associated with each pair of ideas. Materially, this is the Hopfield network equivalent of a semantic space.

3.1.3 Initialization

The initial set of recurrent ideas (including repetitive phrases) extracted from a text comments serves as the input pattern. Each node in the network matching one of the extracted concepts from the text comment is initialized to have a value of 1 (i.e., the node is activated), with the rest deactivated, assigned 0 weight, summarized in Equation 1 below:

$$u_i(t_0) = x_i, \qquad 0 \le i \le n-1 \qquad [1]$$

$u_i(t)$ is the output of node i at time t and x_i indicates the input pattern for node i.

3.1.4 Activation

The nodes contain all important concepts and word combinations from the text. Concurrently, the same network assesses frequencies of occurrence and joint occurrence of different semantic elements within certain structural text units, for example sentences. The resulting graphical structure however does not provide an accurate semantic picture of the analyzed text. An adjustment of the individual statistical weights of the words and relations between them to provide a consistent text representation is needed. The weights of those concepts, strongly related to other frequent concepts in the text should be increased. This is accomplished by assigning the statistical weights of individual concepts to the nodes in the one-dimensional Hopfield with all neurons completely interconnected.

During the activation phase, the statistical weights of relations between concepts are assigned to the links between individual nodes in the semantic network and are permitted to settle. The modified Hopfield network

advances by changing (the weights assigned to the nodes and links between them) to a stable configuration corresponding to the minimum of the Hopfield energy function.

The goal is to obtain the weights w_{ij} once the network reaches a stable state.

$$u_i(t+1) = f_s\left[\sum_{j=0}^{n-1} w_{ij} u_j(t)\right], \quad 0 \le i \le n-1 \qquad [2]$$

f_s is the continuous sigmoid transformation function,

n is the number of nodes in the network, and net_i is the weight computation formula:

$$f_s(net_i) = \frac{1}{1+\exp\left[\frac{-(net_i - \theta_i)}{\theta_0}\right]} \qquad [3]$$

θ_i is the threshold or bias parameter, and θ_0 is modifies the shape of the sigmoid function.

$$net_i = \sum_{j=o}^{n-1} w_{ji} u_j(t), \qquad [4]$$

The term $net_i = \sum_{j=o}^{n-1} w_{ji} u_j(t)$ creates a unique feature of the Hopfield net algorithm in that each activated node computes its new weight (output state) based on the summation of the products of its neighboring nodes' weights and the similarity weight between its neighboring nodes and itself.

3.1.5 Convergence:

Activation continues until the network reaches a stable state, i.e., there is negligible change in the value of the output states between two time steps. To measure the stability of the network, the difference in activation level between two time steps is computed and compared to a predetermined convergence threshold, ε :

$$\sum_{i=0}^{n-1} \left| u_i(t+1) - u_i(t) \right| \le \varepsilon \qquad [5]$$

The parameter ε is a threshold used to indicate whether there is a non-negligible overall difference between two consecutive steps, across all network nodes. The final result is the set of concepts having the highest activation level, when the network converges. The resulting concepts (ideas or noun phrases) are considered most relevant to the concepts contained in the semantic space.

4. Results and Findings

Eighty-one of the 140 respondents gave input to the question regarding suggestions about how to encourage/promote community development activities on the university campus. The total word count for this file is 2,622 words. Text Analyst 2.0 was applied to the text of word-sentences via as described above. Graphically, the key semantic concepts addressed by the respondents are presented in Figure 1 below.

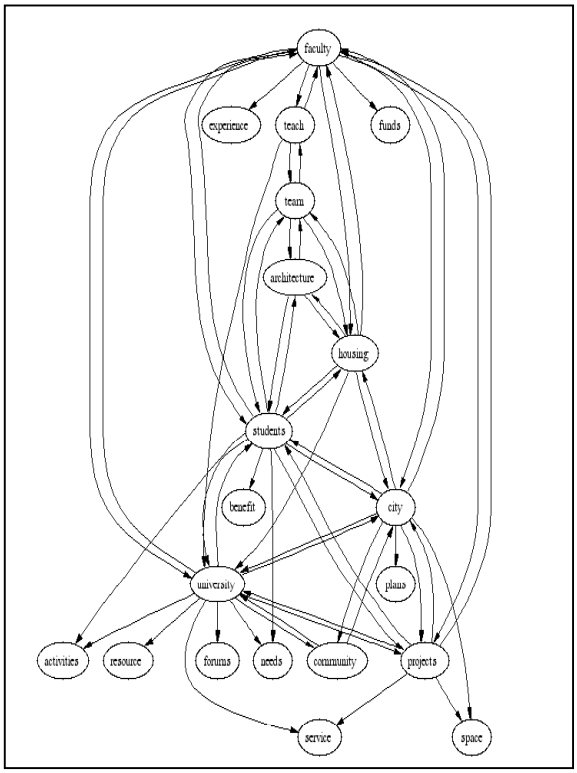

Figure 1: Semantic Map:
Ideation Structure For Encouraging Community Development Activities

The above semantic map presents the key concepts contained in the collective text responses to the question regarding suggestions about encouragement and promotion of community development activities by the university community. It is important to note, this map does not necessarily represent any one person, but

rather, represents the collective or group of respondents. The renormalized weights of words and relations between them are called semantic weights and the resulting reshaped graph-like structure is called a semantic network (which is a list of the most important words and word combinations from the text and relations between them).

A 261-word summary of the total set of text responses is presented below. Note this summary yields highly actionable managerial information:

"The **university** should interact more with the **city** to reach mutually beneficial solutions to employment, **housing** for **faculty** and **students** and community development. Continue to provide a **forum** for discussion of **issues between the university community** and the **city community** at large: Outreach **projects**, which have visible impact. Volunteers who perform **services** for the **needy**, university learning **projects** which enhance quality of life for community, Opening up **university resources** to community. Allow more local businesses to have commercial **activities on university.**

I would be interested in **team teaching** a collaborative course with business, a**rchitecture**, **city** and Regional Planning and Landscape **architecture students** and **Faculty** on investigating **housing** design for families, **elderly** and **students**.

I would like to develop **team**-based senior **projects** (ideally jointly between College of Business, and [local research park] and college of **architecture students**) to **study housing** options for the **elderly**, for instance, and develop viable proposals for **housing** facilities which will fill the gaps in the current market.

We need a lecture series that brings California and Western practitioners to **university** to **share** their **experiences** with the **faculty**.
We need a '**faculty** brown bag' series that allows **faculty** and **students** to come together and **share** local **experiences**

University needs to hold **forums** on relevant topics. Perhaps these initiatives could be tied together with a multidisciplinary **forum** on community development where **university faculty** and staff as well as outside speakers could be invited to **share** their experience and expertise on issues related to community development.

Extended **Studies** is a logical **resource** for promoting **community development activities on our university**."

4.1. Managerial Implications

To encourage community development activities on campus and local community, the content of the survey focused on thoughts, suggestions, and opinions regarding the startup and operation of a community development center at a major U.S. University. The key relational ideas expressed in the group's responses are presented below.

University and city interaction regarding
- **Housing for faculty and students**
- **Services for the needy**
- **Quality of community life**
- **Opening university resources**
- **More local business activities on campus**

Actions within the university to focus on
- **Team teaching**
- **Investigation of housing for the elderly and students populations**
- **Increased relations with local research park**
- **Encouraging practitioners to share experience with faculty and students**
- **Lecture series**
- **Faculty "brown bag" seminars**
- **University sponsored forums**
- **Use of extended studies programs**

References

[1] Bonnet, A. ***Artificial Intelligence: Promise and Performance***, Prentice-Hall International, U.K., 1985.

[2] Dalton, J and A. Deshmane, "Artificial neural networks," ***IEEE Potentials***, 1991, 10 (April, 2): 33-36, 1001.

[3] Chung, Yi-Ming, W. M. Pottenger, and B. R. Schatz. " Automatic Subjective Indexing Using An Associative Neutral Network," ***Association of Computing Machinery, Digital Library***, ACM DL, 1998, 59-68.

[4] Hopfield, J. J. "Neural network and physical systems with collective computational abilities," ***Proceedings of the National Academy of Sciences, USA***, 1982, 79(4):2554-2558.

[5] Knight, K. "Connectionist ideas and algorithms," ***Communications of the ACM***, 33, 1990, (November, 11): 59-74.

[6] Tank, D. W. and J. J. Hopfield. "Collective computation in neuron like circuits," ***Scientific American***, 1987, 257(December 6): 104-114.

[7] Quillian, M .R. "Semantic memory", in ***Semantic Information Processing***, M. Minksy (ed.), 1968, Cambridge, Mass, MIT Press, 227-270.