

A Hybrid Technology Approach to Free-Form Text Data Mining

**Integrating Fuzzy Systems, Self-Organizing Neural Nets and
Rule-Based Knowledge Bases**

A Conceptual Overview

© 1998 Earl Cox



1289 North Fordham Blvd. Suite A312
Chapel Hill, NC 27517

(919) 678-0477

www.scianta.com

A Hybrid Approach To Text Data Mining



You know the appointed end
of all things, and all the ways.
You know how many leaves the earth unfolds in spring,
how many grains of sand are driven by storm and wave
in the rivers and the sea.
You see clear the shape of the future
and what will bring it to pass.

Pindar
Ninth Pythian Ode (to Zeus)
(lines 44-49)

There has been a considerable amount of interest lately in the use of automated tools to derive relationships from large databases. The general techniques, lumped under the ubiquitous umbrella of *data mining* or *knowledge discovery*, have been successfully used in a wide spectrum of industries such as petrochemical, retailing, manufacturing, managed health care, financial services, insurance, and agriculture. Some examples of successful data mining projects include,

- Discovering new bactericides and immune system enhancers
- Finding species specific growth hormones
- Discovering parasite resistant plant strains
- Predicting long range weather patterns
- Product pricing and positioning
- Commercial underwriting risk assessment
- Cross marketing and categorization of customers
- Discovering and exploiting new lines of business
- Discovering potential customers and retaining current customers
- Isolating and controlling etiology of emerging diseases
- Portfolio safety and suitability measures
- Detecting managed health care fraud
- Locating new deposits of petroleum
- Bankruptcy prediction and prevention

These projects all have one factor in common – their data consists of numbers, strings, or symbols arranged in a more or less consistent manner. This consistency is expressed as rows and columns in a relational database, a flat file, or a spreadsheet. In general this allows the data mining analyst to understand a priori the semantics and the structure of the data. Data mining is then concerned with finding, reducing (or intensifying) patterns buried deep in the data.

This fundamental premise about both the structure of the data as well as the predictable semantics of the data is lost when we turn our attention to extracting meaningful patterns from text. The difficulty of mining meaning from arbitrary text ranges from moderate to extreme depending on the organization of the text itself. Articles in scientific journals provide, in most cases, the smallest degree of difficulty since we can exploit such intrinsic structural semaphores as the dependent and independent variables often embedded in the title, the list of keywords, and the jargon of science which carries with it its own semantic relationships. Newspaper and news magazines fall in the middle spectrum of difficulty. Even without the use of textual clues, journal articles, stenographic recordings, newspapers, news magazines, and books have

one significant property that reduces pattern discovery complexity: they are written in a formal, consistent language. This brings us to the most difficult kind of text mining – extracting meaning from free form text. Free form text is often hastily entered by help desk and client services representatives. It most often consists of nonstandard abbreviations, very poor grammar, word juxtapositions, a lack of basic punctuation, a lack of consistent capitalization, and a choppy organization sometimes consisting of individual words rapidly noted and making sense only the “context of the moment.”

In this article we will examine an approach to free form text mining that combines semantic analysis with ambient analytical techniques including fuzzy rule induction and self-organizing (Kohonen) networks. Using this combined approach, the result of actual projects in the insurance and managed health care industries, a fast and effective methodology exists to find, expose and rank important patterns in large databases of free form text.

Text Mining Methods and Problems

As commonly defined in the literature, text data mining involves the extraction of patterns, behaviors, and general knowledge from large collections of textual information. Unlike numeric data, processing text data in any meaningful way involves a large number of considerable difficulties. One aspect of these difficulties arise because textual knowledge extraction demands that we know something about the relationships and meanings of symbolic, semantically rich collections of arbitrary words, images, and mathematical expressions. Another aspect of the difficulties involves the highly unstructured nature of textual information itself. Considerable effort has been directed toward assimilating knowledge from formal documents – journals, books, business forms, wire service messages, etc. – and there are a wide spectrum of approaches that yield reasonable results in this area. However, in analyzing free-form, unrestricted text – transcripts of help-desk calls, telephone messages, e-mail documents and so forth – these same techniques, relying as they do on a standard dictionaries, have not met with nearly as much success. Not only is the text arbitrarily free-form, but, because it is not rigorously organized in any specific context, a high degree of ambiguity and vagueness exists. Consequently an attack on unstructured, free-form text usually adopts a two prong approach: a quantitative analysis of the session properties and a qualitative analysis of the text semantics. Figure 1 illustrates this approach.

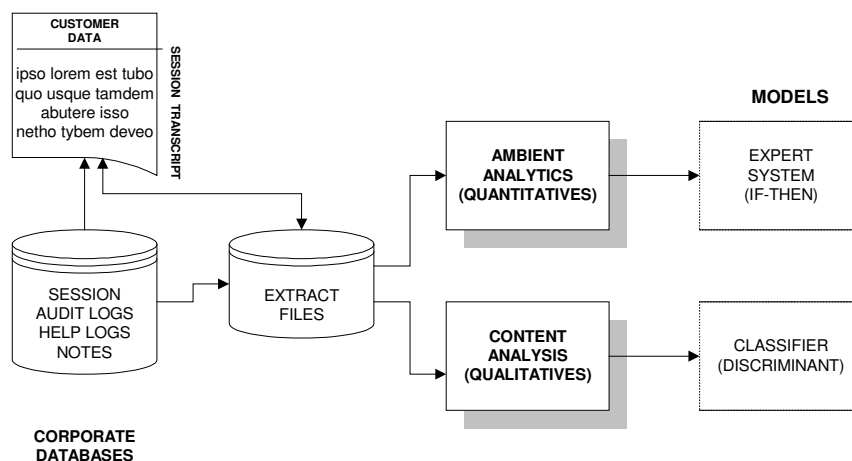


Figure 1. The Approach In Free-Form Text Mining

A Hybrid Approach To Text Data Mining



Ambient analytic analysis, as the name implies, addresses the descriptive parameters surrounding the actual text. In this approach we are evaluating the relationship between retained and lost classes and the generally numerical parameters associated with the audit stream. This technique clusters descriptive properties and produces statistical data points (such as a frequency distributions of calls and length of calls). These are stored in several ways and analyzed by data mining tools that couple fuzzy logic with rule induction. The rule induction facilities generate an if-then rule-base describing how data records are classified into emerging classes throughout the data.

Content analysis examines the text properties of the audit stream. Through a content analysis we endeavor to find patterns in the semantic properties of text that can discriminate between classes of documents. In a customer service problem, the data is scrubbed, noise and background semaphores are removed, tokens and phrases are converted to semantic flags, and the distribution of signal text streams in each document class is calculated. This distribution acts as a filter on the categorization of similar text collections.

Combining ambient analysis with content analysis improves our ability to understand the nature of patterns emerging from large text bases. Building such a system requires the application of technologies from the fields of fuzzy logic, self-organizing neural systems, and semantic rule induction. Figure 2 schematically illustrates how the over-all process is integrated into a cohesive methodology for the discovery and encoding of text knowledge.

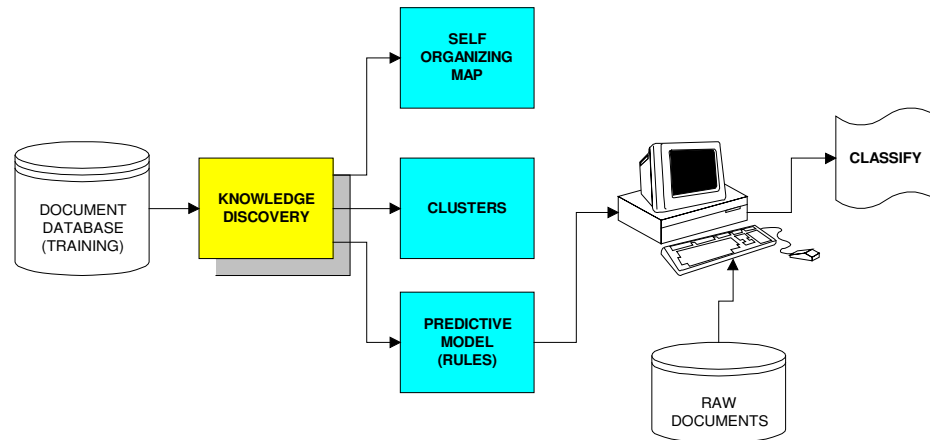


Figure 2. The Text Knowledge Mining Process

Knowledge discovery finds patterns and linkages in the collection of documents. These are reflected in several powerful representational schemes. Self-Organizing Maps (SOM's) are a form of neural networks that bring together related concepts showing their intensity within the database and their proximate relationship with other concepts. An SOM graphical display often uses color coding and indexing to provide a density mapping of the document keywords. Fuzzy Clusters provide a spatial analysis of documents and semantic concepts in the form of related aggregations. We use a form of adaptive fuzzy clustering that determines the proper number of clusters and allows concepts to reside in multiple clusters at the same time (with varying degrees of membership). Finally, fuzzy rules in the form of if-then statements encapsulate the extracted knowledge in a way that can be used to evaluate and classify new documents and make predictions based on document contents. In this next section we discuss, at high level, the two steps used to produce a working text knowledge discovery and modeling system.

Generating Raw Text Vectors

The initial phase of the text mining process, illustrated schematically in Figure 3, involves generating the basic text vectors. A vector represents the collection of non-trivial, noise, and filtered words occurring in the raw text documents. This process reduces the free-form, unstructured documents into a space of semi-structured (but still uncoupled) array objects. These are *n-tuples* (or, perhaps less formally, simply arrays or vectors) containing information about primary text semaphores -- somewhat analogous to keywords and other significant words and phrases, their relationship to the source document, and the type of estimated semantic encoding attached to the semaphore.

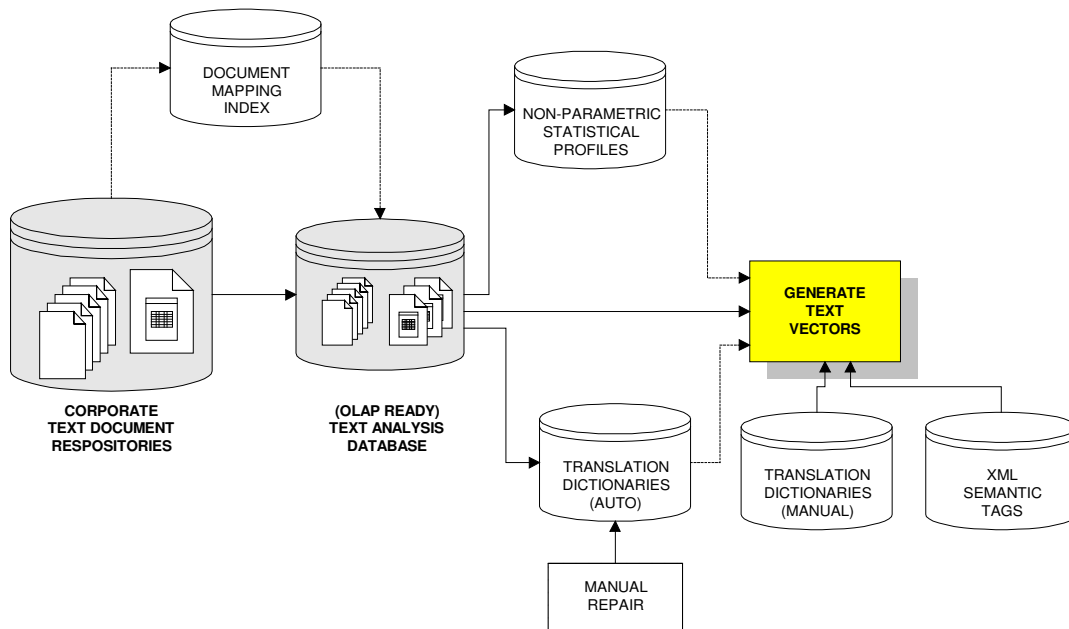


Figure 3. The Text Vector Generation Phase

A text vector in the canonical form $(X_i, D_i, p_{i1}, \dots, p_{in})$ where “X” represents the textual token (or token set), “D” is the document, and “p” are associated properties. The vector generation process sits at the end of a document reduction and synthesis process. This process creates a secondary high performance database of compressed and reduced text objects indexed back to the source. These text objects are prototypical vectors. At the same time we create an analysis database, a data store of statistical information is produced as well as a dictionary of embedded acronyms, abbreviations, and unidentified text strings. Coupled with (optional) a manual dictionary of terminology and an Extensible Markup Language (XML) description of semantic tags, the Vector Generator creates an indexed data store of the vectors.

Quite often the analysis of unstructured textual data presents many difficulties in the mapping of semantic tokens into their proper order and relationships. As an example, a notation in a customer service log for a technical help desk might say:

Calld for hlp on prt docs X21 svr

This is open to at least two related but significantly different interpretations:

Called for help in printing documents using X21 server
Called for help on printer, documents are on X21 server

How we finally resolve the intrinsic ambiguity associated with this annotation depends on several key factors: the amount of contextual information contained in the surrounding text, the expansion or mapping for this token (prt) in the dictionaries, and the frequency of association with printing or printer activities as found and verified in related documents ($D_{1...k}$ associated with the same customer service representative (S_m) as an example). Even so, of course, no mechanical approach can be absolutely precise in collecting and classifying all points of knowledge in a large textual database. The goal of an automated text-mining regime is the fusion of human intelligence (pattern recognition) with machine intelligence (pattern identification).

Semantic Nets

A usual part of the translation of concepts into common semaphores as well as the classification of concepts into taxonomic groups involves the use of semantic nets. A Semantic Network – or simply a *Semantic Net* – is a graph structure representing both concrete and abstract knowledge about a class of problems. Semantic Nets are problem dependent, that is, the interpretation and exploitation of semantic knowledge is based of the model state and context. We generally design and construct a semantic net to represent the knowledge in a specific problem, although some general concepts, of course, can be shared among many widely diverse models.

A Semantic Net describes the relationship between exemplars, concepts, features, and processes in the text-mining model. By “semantic” we mean that the net imparts a meaning to the relationship between two or more nodes in the system. The network itself consists of entities of concepts connected by the edges of a graph. It is the edges that carry the relationships in the semantic net. Figure 4 shows the basic architecture of a semantic network.

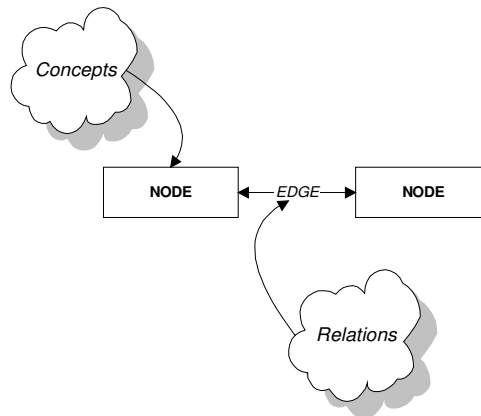


Figure 4. The Structure of a Semantic Net

The entity at the node can be an object (a physical ‘thing’), a concept (an ‘idea’ or an abstraction), a feature (a property of the network), or a process (a property of the interaction between two elements in the network, which can themselves be processes). In general, we refer to a non-edge element in a semantic

network as a *node*. Thus, we can say that node **X** is related in some way to node **Y**. In the simplest form, the relationship indicates that **X** is a type of **Y**. Figure 5, as an example, says that a car is a type of vehicle.

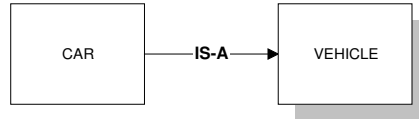


Figure 5. A very simple semantic Net

The relationship “IS-A” makes this type of relationship explicit. We note that the relationship exists in the edge of the network – that is, it is a property of the network, not a property of the node. In Figure 6, both nodes are concepts – one more general than the other. Although we specify a semantic net in terms of explicit operators, the net manager constructs a bi-directional linkage for each concept. This, if we define *car is a vehicle*, then we also have, as shown in Figure 6, a backward edge operator.

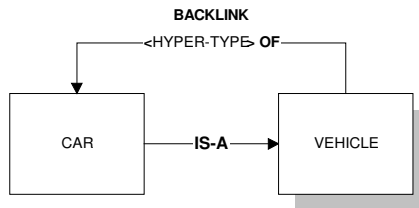


Figure 6. The Backward HyperType Link

Such backward links, called HyperTypes, are automatically maintained by the semantic net handler and establish the principle method of walking forward and backward through the semantic graph. Backward chaining through the semantic net provides a method of focusing on components elements in complex concepts – thus we can find the various kinds of exemplars that make of a class of objects called a vehicle (such as cars, trucks, buses, motorcycles, and so forth.)

It is this association of text elements with their semantic classes that is carried forward into the final text vectors and provides the self-organizing map as well as the fuzzy clustering facilities with their root ability to find patterns and associations in the data.

Building Diagnostic and Predictive Modes

The canonical vectors provide the raw material for knowledge discovery and are used to generate a wide variety of information rich models. Figure 7 provides a conceptual overview of how the vector store, coupled with both the statistical data and the XML semantic data, is used to produce a set of related analysis tools.

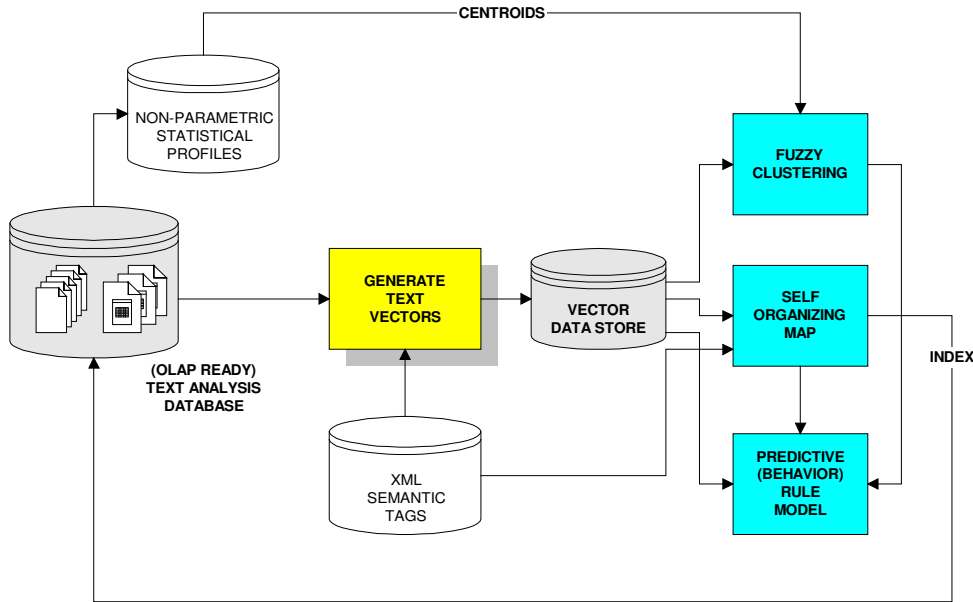


Figure 7. Producing the Analysis and Prediction Models

Like conventional numeric data mining, text mining is concerned with uncovering deeply buried patterns in the data. These patterns are often tied to outcome states in the mind of the client. That is, in a customer service application, we might like to know which representatives are the best (or worst), what problems occur most frequently (and how are they handled), what are the most difficult problems (and how are they handled), is there a relationship between customer satisfaction and problem handling or resolution, do particular departments, demographic regions, or clients have unusual problem reporting and resolution patterns?

These problems are solved in a hybrid text mining approach that combines semantic analysis with an analysis of the statistics associated with both the text and the context of the text. As an example, in one insurance industry project we might like to find those customers who are likely to drop their policy. Is there a significant relationship between their interaction with customer representatives and their failure to renew their policy? Fuzzy rule induction, as one leg of the hybrid architecture, isolates patterns based on the frequency of keywords, the density or frequency of semantic terms, the association with service representatives, the number of words in the text, the length of the interaction, the time of day, as well as customer background and demographics (sex, age, length of time as a policy holder, residence location, approximate annual income, and similar statistics). A rule induction process generates fuzzy rules that classify clients into two categories --C1 (retained) and C2 (dropped) -- with their predicted membership in

A Hybrid Approach To Text Data Mining



either (or both) of the categories. In this way semantic patterns and analytical patterns are brought together to sieve through large free form text database is a directed and powerful search.

The hybrid approach has many technical as well as client advantages. In this mixture of traditional text analysis based on semantics and analytical analysis based on fuzzy rule induction and self-organizing maps we can balance the objectives of a project to focus on which ever technologies provide the best insight into the data. Thus we follow the old dictum of playing to our strengths. In particular, apply rule induction and self-organizing pattern discovery to the actual semantic patterns themselves (in the form of both the semantic net relationships as well as XML tag vectors) gives us a deep insight into the robustness and structure of the actual information emerging from the text. In projects ranging from insurance policy management, health care fraud detection, and breast cancer treatment the combined analysis o produced highly significant results with low error rates and very low false positives.

For more information or to schedule a presentation call (919) 678-0477 or visit www.scianta.com



©2004 Scianta Intelligence, LLC
AR-PA-007