Text Mining: The New Data Mining Frontier

Traditional Approach to Data Mining

Most experienced data mining practitioners are familiar with mining data when the data presents itself in a structured format. A structured format means that the actual records themselves are assembled in rows while the information that is used for data mining resides in columns.

Records represent the level of detail on how information is being captured. For example, records can be at the customer level, transaction level or at any level where information is being captured. Columns running down the rows of data can be thought of as variables that depict a certain piece (field) of information pertaining to that record.

Chart 1 and 2 below are examples of some structured data formats:

Table 1

Customer No.	Household Size	Postal Code	Income
0001	3	L1A3V1	125000
0002	2	M5S2G1	30000
003	1	H4B2E5	40000

Table 2

Transaction No.	Date	Amount	Product Type
000001	July15/2009	100	A
00002	Oct.1/2009	75	A
00003	Sept.15/2009	200	C

Chart 1 is an example of a Customer file and Chart 2 is a Transaction file. In the Customer file, individual customers are presented in rows while customer number, household size, postal code, and income are variables presented in each column. In the Transaction file, individual transactions are the rows while transaction number, date, amount, and product type are different variables presented in each column. This structured approach provides great flexibility when it comes to data manipulation and the ability to derive new information from source data. For data miners, this is a critical capability because of the potential it offers in deriving new insights from existing information.

Text Mining

More recent advancements in the field of data mining now focus on the notion of “text mining”. Text mining is similar to the standard knowledge discovery process of mining structured data, except that in text mining the analyst is working with unstructured data.

Unstructured data is textual type data such as email, phone conversations, open-ended responses to surveys and conversations within various social media applications. Table 3 below contains a very simple example of what an unstructured format might look like.

Table 3

Customer No.	Email
0001	I really like the RRSP product and will continue to invest every year. However, the level of service is sub-standard and I will be looking at other companies. But it will be difficult since I have so many products with this institution.
0002	I wish this company would be more proactive in offering me products and services that truly meet my needs. The customer service people are great and are simply doing their job. But the company is clearly not thinking of my real needs.
0003	The level of service is outstanding and I am extremely interested in hearing about all your products and services. Could you send me more information on them?

In this example we have three customers, each have sent one email to the company. The data mining task here would be to analyze the information listed under the column heading entitled “Email”. The analyst would be trying to derive information from the Email column. But how? In the structured data world for example, an analyst can easily create a yes/no variable defining whether or not a person lives in Quebec based on the postal code field of where a person lives. How can the analyst create binary or yes/no type variables based on some condition, create change type variables, or mathematically derived variables such as mean, standard deviation, median, etc. in the unstructured data world? The challenge is to derive new variables from a series of sentences. The critical point to remember is that there is information with the text that can and should be used.

The process of using this information by creating variables from unstructured data is not necessarily more complex than the process of creating variables from structured data. It is just different a different process that requires a different and a different set of data mining tools. What follows provides an overview of this process.

The Text Mining Process

The first step in trying to mine unstructured text data is to perform some clean-up or data hygiene. This step involves the elimination of text that provides no meaningful information. Examples of this include punctuation such as period, comma, etc. as well as prepositions such as “the”, “of”, “and”, etc., and pronouns such as “she”, “he”, etc. From the text in Table 3 above, data hygiene techniques might parse and reduce the content of three emails to the following:

Customer No.	Email
0001	really like RRSP product continue invest every year level service sub-standard looking other companies difficult have many products institution
0002	wish company proactive offering products services meet needs customer service people are great are doing job company clearly not thinking real needs
0003	level service is outstanding am extremely interested hearing products services send information

After excluding extraneous text, we have attempted to retain only the key information that we need to extract meaningful information.

The next step in the process is to perform a frequency distribution on all the words that remain from the data cleansing process. Here we begin to obtain our first glimpse of meaningful information. We can see how often certain key words occur in the text. Let's go back to our example and take a look at what this might look like after a simple frequency distribution:

Table 4

Words	Frequency
Products	3
Service	3
Company	2
Level	2
Needs	2
Services	2
RRSP	1
Am	1
Are	1
Clearly	1
Companies	1
…. Etc	

From the above information, we can see that the discussion among this very small sample of three customers tends to focus on "products", "service", "company", "level" and "needs".
While this finding provides some insight directionally, it is premature to enable us to draw specific conclusions. The next and arguably trickiest stage of the process is required to do this.

This third major stage in the text mining process attempts to find relationships between words and phrases that seem most prominent in the text. The linking of words together and creation of common phrases represents the key selling features for most vendor companies that sell text mining software. This is particularly true for text mining vendors selling to government agencies that may be trying to identify criminal activity through the analysis of phone conversations.

Two of the more common statistical techniques that could be employed to identify words that tend to occur at the same time include Correlation analysis (keyword analysis) and Cluster or K-means analysis. Correlation analysis is used to find groups of words that

appear together most frequently. K-means or Cluster analysis attempts to find phrases or groups of words that appear as common themes. At the same time, the clustering technique is attempting to identify groups of themes that are different from one another. This is somewhat similar to the traditional use of clustering to create distinct customer segments where customers within each segment have similar behavioural and demographic characteristics.

In our three record example, some emerging themes might be:

1) Wide Range of Products and Services
2) Need For Information
3) Service Levels

Sentiment Analysis

But even as we identified the three common themes that seem to emerge from these records, the notion of how the customer feels or the sentiment that he or she is trying to convey is certainly not evident here. The ability to understand the sentiment or emotion behind the statement is often referred to as “sentiment analysis”, and is an area of extensive research within the text mining discipline.

For example, one could have statements such as:

1) “I am really going to like this company now that they have cut services and it takes me longer to get through to a live agent”

or

2) “I am really going to like this company now they have invested in new customer service software and hired another 1,000 agents”

One can see that the first statement is sarcastic and expresses a negative sentiment while the second statement appears to be quite positive. The ability to detect the specific sentiment in each of these statements can easily be understood by the human brain based on our understanding of language. But it becomes much trickier to design software that can decipher the sentiment belying these two statements. In fact you might have two statements that both state “I really like your company” and only the tone of how these statements are uttered would indicate the true sentiment. Text mining tools might never pick up these differences in tonality, but an offshoot of text mining called “speech analytics” that analyzes unstructured data but in voice form, might indeed offer this capability. Analyzing sentiment within voice type data is also being extensively explored among speech analytics vendors.

The Marketing Opportunity

Text mining presents an opportunity for marketers to understand customer engagement in an unstructured data environment. As the volume of interactive unstructured digital data grows, Text Mining will become increasingly important as a way to obtain customer insights, allow marketers to classify customers based on sentiment (negative, positive, or neutral) and classify customers into theme categories based on their individual discussions.