White Paper

Beyond Search-and-Retrieval: Enterprise Text Mining with SAS®

by Stephen E. Arnold
ArnoldIT.com

Table of Contents

Enterprise search and its laundry lists of results are increasingly frustrating to users of enterprise search systems. What can an organization do to break through the log jam of often-irrelevant results and "missing" information? How can an organization make "finding information" reduce costs instead of increasing them? What can you do to squeeze more value from the digital data and electronic documents you have? SAS has an answer: text analytics.

1. Enhancing the usefulness of existing information in an enterprise

a. Integrate discovery in databases with unstructured text

Enterprise search is undergoing an important change. Systems using brute force matching of the words in a user's query are becoming more pliant and discovery-centric. Text mining is coming to an enterprise search system near you. Its arrival is long overdue. A system that can seamlessly integrate discovery in databases and unstructured text can pay huge dividends quickly. (See Stephen Arnold's article in the November 2006 issue of *CXO America*: http:/www.cxoamerica.com/pastissue/printarticle.asp?art=25408.)

Many CFOs, CEOs, CIOs and CTOs (hereinafter, CxOs) have learned the painful way that in their organizations, text-centric information problems are common – and expensive to address using traditional management techniques such as hiring temporary staff, outsourcing or trying to get existing information technology to handle customer support, marketing and legal tasks. This analysis will outline the emerging topic of text mining as a potential answer to reduce the dissatisfaction with existing manual processes. Text mining provides a way to enhance the usefulness of existing information in an enterprise. By adding value to its information, a company can help everyone in the organization work smarter. Data mining is the process of data selection, exploration and building models using vast data stores to uncover previously unknown patterns. What does this mean to you? You can produce new knowledge to better inform decision makers before they act. Build a model of the real world based on data collected from a variety of sources, including corporate transactions, customer histories and demographics, even external sources such as credit bureaus. Then use this model to produce patterns in the information that can support decision making and predict new business opportunities. Text mining capabilities enable you to apply such analyses to text-based documents. With SAS' rich suite of text processing and analysis tools, you can uncover underlying themes or concepts contained in large document collections, group documents into topical clusters, classify documents into predefined categories and integrate text data with structured data for enriched predictive modeling endeavors.

Whether it's used to drive new business, reduce costs or gain the competitive edge, data mining (and its close cousin text mining) is also a valuable asset for organizations. Especially in the area of customer relationship management (CRM), applying the techniques of data mining has become an accepted "best practice" for companies serious about business. No longer can successful organizations rest on their laurels and continue to do business as they always have in the past. Only those that take proactive measures will survive in today's competitive world.

Text mining, however, is not well-known because it is viewed as an emerging technology with untapped potential by those eager to analyze the textual-based data that is accumulating and filling their vast data warehouses. Enterprise search vendors have asserted that their keyword matching systems are, in effect, text mining systems. Because text mining is not well understood, these assertions lead many organizations to believe that broad categories of results equals text mining. This is not true and the situation can cost an organization money and opportunities. Enterprise search that generates a long list of possible matches wastes the user's time. Every minute spent browsing and clicking on "hits" costs the organization money and reduces the employee's ability to respond to an information need quickly. Tools, like those available from SAS, provide alternative ways to find an answer, pinpoint the fact needed, or get 'oversight' and insight into structured and unstructured data.

Organizations ought to take steps and begin to harness more of the data they are adding to their growing data stores – structured and unstructured alike – so that they will be empowered to take proactive steps and improve decisions to guide their companies to a brighter (more profitable) future. Text mining software can surface patterns and trends that you may never have thought to look for. With comprehensive data mining solutions such as SAS Text Miner, you can go from raw data to accurate, business-driven analytical models in a seamless, efficient process.

b. Difference between search and discovery

It's important to remember that search engines are designed to return a specific subset of documents based on a query performed by the user. Typically, users know what they seek and construct a query to accomplish the goal. The search engine then returns a subset of documents that matches that query.

In contrast, text mining engines are designed to help the user discover key concepts within documents or subsets of documents without having to read the entire collection. Unlike the user of a search engine, text miners are unsure of what they want to know and are relying on the collection as a whole to speak for itself.

Text mining software is designed to process unstructured information. SAS has had success in selling its technology where customers must find meaning and trends in e-mail, call center feedback from customers and textual information from individuals who file reports in electronic form. SAS Text Miner can be used to discover common themes in collections of unstructured information.

SAS Text Miner can be optimally integrated within your existing IT infrastructure with a single, unified system. The SAS Enterprise Intelligence Platform transcends organizational silos, diverse computing platforms and niche tools to deliver new insights that drive value for organizations. SAS customers have ready capabilities for data integration, analytics and business intelligence (BI). This ensures data can be cleaned, manipulated, prepared and fed into analytics for an online analytical processing (OLAP) engine, forecasting or other analytical modules before displaying results via Web graphics or interactive flexible business intelligence (hereinafter, BI) tools.

c. How SAS® Text Miner works as a discovery search tool

SAS Text Miner technology delivers extraction, automatic classification and discovery to SAS customers. The outputs of the SAS Text Miner subsystem allow you to view term statistics, identify similar documents, and view term and concept relationships in a graphic display. The system also performs automatic classification and meta tagging.

If you have an existing taxonomy, SAS Text Miner can use that data and discover new categories. The system can identify clusters of related information such as call center comments or statements within e-mail. Stephen E. Arnold, ArnoldIT.com, calls this "information oversight" which, he says, "Makes it possible to spot a key fact or the needed information without wasting time clicking on links and hunting for the item of information."

The data generated by SAS Text Miner can be used to enhance an existing search system. However, this data can be used for predictive modeling by itself. Alternatively, you can merge discovered data with existing structured data and perform

d. How SAS® Text Miner enhances the discovery process

SAS' text analytics are an add-on to its premier data mining software solution – SAS Enterprise Miner™. The SAS Text Miner product is surfaced as an additional node to fit comfortably within the predictive analytics umbrella. SAS is the leader in business intelligence and predictive analytics software. SAS Predictive analytics software provides a wide range of integrated capabilities for time series analysis and forecasting, econometrics and systems modeling, and financial analysis and reporting, with direct access to internal/external commercial data warehouses/marts. With 31 years experience and 43,000 customer sites worldwide, SAS helps manage performance by transforming data into predictive insights.

SAS' system assumes that the primary users will have strong understanding of statistics and basic navigation familiarity with the Java GUI of SAS Enterprise Miner.

The benefit to the SAS Text Miner user is that virtually all of the natural language processing and statistical functionality is available within the SAS graphical environment. SAS has done the heavy lifting required to make algorithms available as drag-and-drop or point-and-click functions. According to Mary Crissey, Analytics Product Marketing Manager at SAS, "A key benefit to the user of SAS Text Miner is that the time required to graft linguistic functions onto a model is reduced by more than 90 percent."

Using intelligent algorithms and lexical processing to automate the categorization, tagging, or building of topics and document indexes:

- Provides automatic identification and indexing of concepts within the text.
- Visually presents a high-level view of the entire scope of the data processed by the system, with the ability to drill down to relevant detail.
- Enables users to make new associations and relationships and to present paths and links for further document analysis.

In 2006, SAS moved quickly to become a Google partner. Google OneBox API allows SAS to offer the Google search interface to core SAS data and reports. In addition, as the Google Search Appliance becomes more robust, SAS has the opportunity to offer collaboration and access to other Google functionalities. It's too early in the adoption cycle to be able to quantify the value of the Google relationship, but as a privately held software company, SAS is able to move quickly and decisively.

In the same way that SAS has expanded access to BI by creating targeted user interfaces for its software that match the skill levels of individual users, SAS and Google will provide joint customers who activate the new Google OneBox for Enterprise feature of the Google Search Appliance with a familiar, secure way to search for real-time information delivered by SAS Enterprise BI Server software. In addition, the combination of the Google Search Appliance with the SAS Enterprise Intelligence Platform will give users more information than ordinary keyword searches can provide. SAS' contextually relevant search capabilities with Google not only explore metadata but also look at the business views (SAS Information Maps) that have been defined by SAS clients.

The vision ahead for SAS will not be limited to its premier data analysis (statistical processing) capabilities. Not only is SAS gaining foothold in query and reporting software (BI), it in fact extends across the entire SAS Enterprise Intelligence Platform as a complete end-to-end system that includes the data warehousing ETL which stands for extract, transform, and load (handling data input issues). The potential for SAS to deliver THE POWER TO KNOW® seems unlimited because SAS is meeting the challenge for small and large databases stuffed with traditional structured data that can be measured and quantified as well as the qualitative descriptors of data referred to as "unstructured" – text-based content that can occupy gigabytes of information in organizations. Examples of traditional structured data are age, job classification and income information. Examples of unstructured data include text-based content in Web pages, e-mail messages, articles, memos, customer feedback, warranty claims, patent information, surveys, research studies, resumes, client notes, competitive intelligence and more.

2. SAS® Text Miner is part of the SAS® Enterprise Intelligence Platform

a. Overview

The SAS platform optimally integrates individual technology components within your existing IT infrastructure into a single, unified system. The result is an information flow that transcends organizational silos, diverse computing platforms and niche tools – and delivers new insights that drive value for your organization.

The core notion of SAS is that analytic processes work off a single version of the truth. By defining an analytical platform and setting up a metadata (information about data sources, including how it was derived, business rules and access authorizations) architecture, SAS provides a centralized repository for data and information. SAS has created a variety of interfaces and tools to permit third-party applications to access information managed by SAS. SAS' principal strength is that a client no longer has to bear the burden of figuring the most cost-effective way to manage the complexities of integrating different repositories of data.

The company added text mining to its technology suite six years ago when interest in unstructured text first rippled through the analytics sector. SAS' approach was to license modules from Inxight for stemming, part-of-speech tagging and entity extraction technologies so they could link with the clustering and pattern recognition facilities of the SAS text mining solution. By contractual right, SAS continues the use of Inxight technology as it did before the Business Objects acquisition of Inxight. SAS Text Miner will continue to offer advanced technology for the integration of structured and unstructured data. Additional solutions focused to industry needs are already under development at SAS. Any future potential replacement of modules will be integrated into new releases of SAS technology which current customers get at no extra cost.

SAS' approach is designed to provide a single environment for data analysis, text mining and knowledge discovery. SAS' marketing team likes to say, "SAS gives customers the power to know." SAS Text Miner has tightly integrated text-based information with structured data. From a single interface, an analyst can obtain a complete view of structured and unstructured information to enhance analyses and decision making.

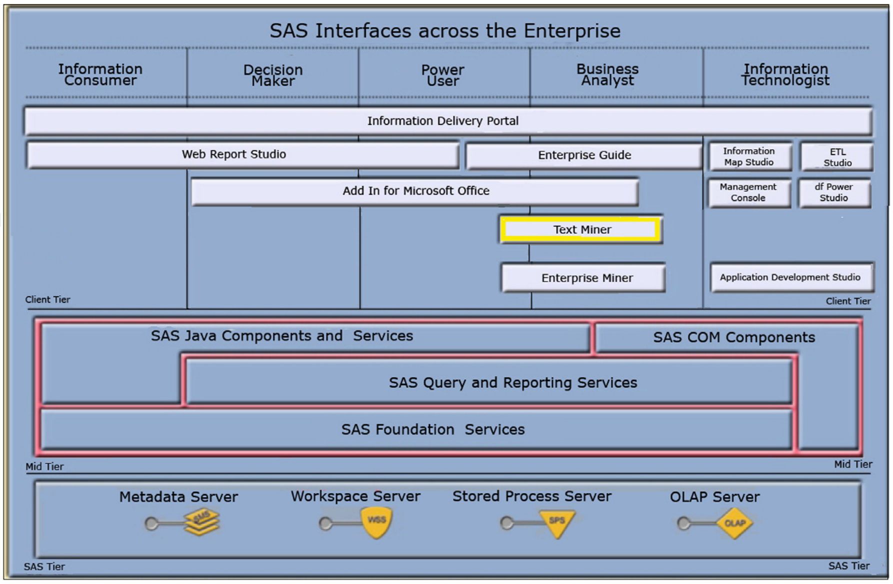
With SAS, no third-party tools are required, and SAS can "play nicely" with most third-party systems because it accommodates data in a myriad of formats and running on a wide variety of operating systems. If a proprietary function is needed, SAS permits unlimited customization. Configuration assistance and installation experts are on hand to guide the customer's SAS administrator in the role of managing the servers and software that comprise the SAS framework.

SAS' text mining subsystem is available as an enterprise application (installed on a server and linked to many client computers) or a quasi-standalone subsystem (installed on a single laptop). Both text mining subsystems can:

- Filter e-mail for regulatory compliance or security applications.
- Group documents into predefined categories.
- Route news items.
- Cluster analysis of documents in a collection, survey data, review customer complaints and comments.
- Predict financial shifts from business information and customer satisfaction from customer comments

b. SAS® creates an information circuit for organizations

SAS is a system, not a single software package. The figure below shows the overall SAS architecture. The SAS Text Miner module (highlighted in yellow) sits in the client tier of the SAS framework. SAS Text Miner is a subcomponent of the SAS Enterprise Miner subsystem.



SAS, now in V ersion 9, provides a comprehensive informaiton environment. It includes servers, data management tools, analytic routines, APIs, and such advanced features as the exchange of data, scipts, andmodels across the dozens of SAS modules. Featuring full support of portal accessible text mining, SAS provides a comprehensive solution to organizations that need to analyze large volumes of textual and numeric data.

In this information circuit, information flows into the various servers. Analysts manipulate the data in the servers using a range of tools. Users of the data can access the reports in standard office software applications such as Microsoft Excel.

Before looking at the specific functions of the SAS Text Miner module, let’s review SAS’ twin development paths, which have yielded some important innovations.

The first is the development of the SAS Intelligence Storage system. The SAS customer can rely on SAS’ framework to provide data warehousing, integrity analysis tools and administrative interfaces when manipulating the gigabyte and terabyte collections of data found in many organizations. For the analyst, the data management techniques do not intrude on the specific statistical processes an individual analyst uses. However, those accessing and manipulating data see that the speed with which SAS handles certain data intensive processes has increased.

To illustrate SAS’ storage efficiency, consider an Oracle table containing 10 terabytes of XML data. The Oracle database will require 10 terabytes for the table data and another 30 to 40 terabytes of storage for indices and the swap space that will allow for efficient data manipulation. The cost of storage is dropping, but the volume of data is increasing. It makes sense to minimize disc space requirements in order to hold down costs and to have the ability to process new data. The SAS Intelligence Storage system requires 10 terabytes for the data, but SAS requires an additional 10 terabytes of space for metadata and data manipulation. SAS has mostly removed the bottlenecks while implementing a multithreaded kernel capable of scaling up to handle clusters of commodity servers. At the same time, SAS has reduced the burden on the database administrator.

The second innovation has been the refinement of tools for individual analysts and the services available to an analyst to create a boiled-down, graphical snapshot of key findings. If the tools built into the SAS enterprise mining and text mining subsystems are not adequate, an analyst can turn to the company's graph product and create Madison Avenue-style charts complete with wizards and previews of a graph or chart.

The components for SAS Text Miner are the SAS® 9.1 system, SAS Enterprise Miner, Java, necessary hardware servers. Before SAS Text Miner can be used, SAS requires that a basic statistical foundation be installed. SAS Enterprise Miner installs on that base. Finally, the SAS Text Miner code installs into the SAS Enterprise Miner module. There is, therefore, no way to use the SAS Text Miner without licensing other software from SAS. Pricing can vary depending on the specific requirements of the client organization.

The platform concept is an important one. The idea is that a licensee can build an enterprisewide information analytics system on SAS technology. SAS' descriptions of its platform simplify to a great degree the large number of components the company has developed. For example, a licensee can acquire software that facilitates grid computing. SAS also offers a storage management system. Both grid and storage technology are pivotal to the success of a large-scale text mining application.

3. SAS – The company

a. History

SAS can boast that it is one of the largest privately held software companies in the world. Headquartered in Cary, North Carolina, the company generates about US$1.5 billion in annual revenue. The company's analytic products range from statistical toolkits to enterprisewide data analysis systems. The company's software is included as part of the curriculum in data mining and BI in the US, Europe, Australia and elsewhere.

SAS began at North Carolina State University in 1976, when Jim Goodnight, now CEO, launched the new company with three of his colleagues. The company has grown to encompass 10,000 employees and 400 offices worldwide.

Dabblers in analytics will find SAS' products inscrutable. Training in specific SAS technical approaches is being incorporated into more and more university campuses today – especially into business programs, which value the competitive advantage analytics provides. SAS public training centers are located around the world, staffed by certified instructors who also travel to customer sites as well as teach via Live Web. SAS Self-Paced e-Learning tutorials are becoming quite popular. SAS Press publishes a wide variety of books that clarify the terminology, technical framework and best practices customers are implementing with SAS software. Software user manuals and documentation are posted on the Internet for the public to see, so even those who have not yet installed the software can look up the mathematical algorithms incorporated in a certain procedure or product of interest.

SAS has worked diligently to make its tools accessible and user friendly in response to market demands and feedback from its large customer base across the globe. Today's users attracted to the data mining solutions include traditional statisticians, business managers and university researchers as well as domain-specific industry professionals.

Today SAS is recognized as a software provider that does more than number-crunching and fancy data manipulation. Analysts have stated that SAS sets the standard for BI.[1]

[1] http://www.sas.com/news/analysts/by_technology_bi.html.

For those unfamiliar with SAS, the company provides several choices for kicking off numerical analysis processes, such as the SAS Enterprise Guide point-and-click interface or JMP® software (version 7, released in May 2007, links to SAS data sets). Stored services now provide automatic ways of rerunning common tasks so programming tools can be saved and quickly automated. These built-in options and user interface make it easy for business managers to produce and understand reports generated by SAS. Programmers trained to use SAS can make jumps through flaming hoops by adding their own creative touches with personalized macro programming or tailored interfaces. SAS wants to enable its customers to deploy as basic or as complex a text mining system as they desire. Then, when it becomes necessary to add additional functionality – for example, advanced data analytics – SAS components will work smoothly.

SAS continues to maintain its leadership role in the analytics market as it has been doing for more than three decades. New niche vendors enter the analytics market sector, but none possesses the breadth and depth of SAS analytics completed by strong data integration and BI capabilities across the platform.

The programmer-analyst with a strong foundation in statistics will revel in SAS' technology and its power. A person who invests the time to learn the SAS approach will find that SAS delivers usable, flexible and scalable analytic tools.

b. Customers

Managers and professionals with SAS experience are strong advocates of the system. In articles and speeches, SAS customers emphasize:

- Reduced time-to-decisions and a more accurate organizational view. By combining structured data and unstructured text, and automating the process of analyzing the data, SAS Text Miner helps organizations gain meaningful insights that successfully drive overall business direction.
- Improved organizational performance. While most software vendors offer classification of one text field into a single class, SAS Text Miner enables the classification of multiple structured and unstructured fields. It improves your organization's performance by distilling information from multiple business units into a format that is easy to manage and analyze.
- Ability to recognize trends and predict business opportunities. Analysis of information such as customer letters and call center notes may provide valuable information about customer dissatisfaction or insights into service and product needs.

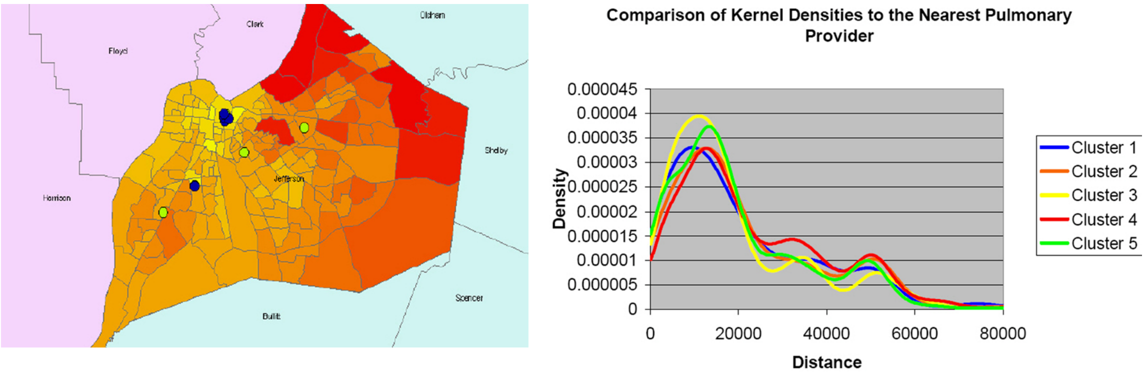
SAS has implemented a number of interesting text mining systems for clients worldwide. Two examples below illustrate the use of the SAS text mining system in both academia and business:

(i) University of Louisville uses text mining, data analysis and geospatial mapping to understand medical conditions and improve patient care in hospitals.

(ii) Honda uses text mining in warranty and quality analyses.

(i) Understanding medical trends

The University of Louisville used SAS text and data mining technology to identify major trends in several cardiac and respiratory conditions, and to analyze correlation such as median income and geographical locations for Jefferson County patients in Kentucky. SAS and ArcGIS were used to compile, author, analyze, map, and publish geographic information and knowledge.



SAS in combination with ESRI tools generated graphic displays of data on maps of Jefferson County and kernel density comparison graphs of the combined data from structured tables and text mining.

This study ultimately demonstrated that text mining and clustering allowed for efficient statistical data analysis. Furthermore, the bridge between SAS and ESRI allowed for a statistical formulation of the data as well as a visual interpretation of the patients and healthcare provider locations in Jefferson County.

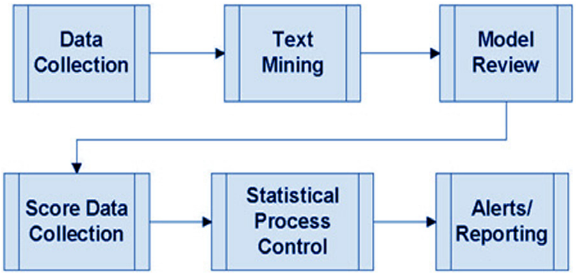
According to author Patricia Cerrito:
"SAS has integrated its analytic tools better than any other software vendor. Text Miner highlights relevant patterns in documents such as clinical reports, and it quantifies text-based information. We move seamlessly to Enterprise Miner to combine and analyze this unstructured text with structured data such as demographics and laboratory values. And, we augment that analysis with SAS/STAT That's why we standardized on SAS for our research. No other software delivered this depth and breadth of analytic functionality. In addition to superior algorithms, SAS delivered the simplest interface for managing and importing data from any source. During any given work day, these benefits add up to significant ROI at the University of Louisville."[2]

(ii) *Improving quality in automotive manufacturing*

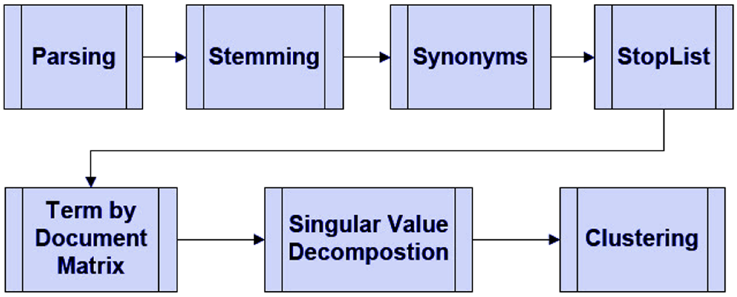
American Honda launched an initiative to extract early warning of potential product problems from customer feedback. An alert system was implemented that used SAS Text Miner and SAS' quality control module. Although analysts already used comment fields to understand automotive problems under investigation, the new system has to "read" incoming freeform text data systematically and generate outputs that could be analyzed by automated SAS Analytics. The system used clustering models to group similar warranty claims together. New data about that particular auto model was then processed, scored and placed into the best cluster. American Honda analysts monitor cluster sizes and variances weekly. Deviations from expected values are flagged for human review.

The overall Honda process consisted of taking warranty and customer data (both structured and unstructured) and creating a common SAS representation. The data moved through a model review, scoring and statistical process control sequence. The outputs generated the alerts when certain warranty-related values were above the expected threshold.

[2] Dr. Patricia Cerrito, University of Louisville, quoted by SAS in a December 2003 news release. See http://www.sas.com/news/preleases/121503/news1.html



The text mining process highlighted in the figure above consisted of seven distinct steps. The outputs of these steps were clustered items.



4. Data mining products available from SAS

a. Highlighting the SAS® Text Miner solution

SAS Text Miner is an add-on to the SAS Enterprise Miner application. SAS Text Miner runs as a node within the larger SAS framework. SAS Text Miner provides a number of tools for discovering and extracting knowledge from text documents. The system transforms unstructured textual data into a usable, structured format. The unstructured data feeds into the SAS System which transforms that into documents that are classified, displayed in relationship diagrams, clustered into categories and/or incorporated with other structured data for further analysis. The outputs of SAS Text Miner can be counted, parsed and analyzed in several ways.

SAS Text Miner provides tools for discovering and extracting information from a widely varied collection of text documents. The tools can uncover verbal patterns and concepts that are embedded within the e-document collection.

i. Textual data comes in many flavors

SAS Text Miner can process volumes of textual data including:

- Web content.
- E-mail messages.
- News articles.
- Research papers.
- Surveys.
- More than 200 office file types.

ii. Competitive differentiators

SAS' approach to text mining has three distinctive features:

- Text mining is tightly integrated into SAS' broader system for statistical analysis. The benefit is that an analyst can use tools as opposed to having to figure out how to get different tools to "play well" with other applications.
- The programming and procedures expertise of SAS eliminates most of the learning curve associated with text mining.
- Data from the text mining module – what SAS calls a node – is available to other SAS processes and procedures. Data from text mining can be merged with numeric data from other analytic processes without a formatting step.

Because SAS Text Miner runs as a function within the SAS operating system, its outputs can be further processed without conversion by any other SAS software module. This integration of text mining outputs into usable formats is a definite advantage. SAS was one of the first companies to deliver a text mining solution able to merge existing structured data with the structured outputs of its text mining subsystem. By combining text and structured data, the reliability of predictive analyses improves.

iii. Key features

SAS Text Miner guarantees compatibility with other SAS components. Furthermore, the text mining subsystem can be integrated into the SAS platform without custom programming. Other features of SAS Text Miner include:

- An integrated interface for analyzing text (unstructured data) in conjunction with multiple related database (structured) fields. The graphical user interface significantly reduces text mining time for both business analysts and statisticians.
- Text parsing to extract terms or phrases from large text collections, while stemming terms to root form based on parts of speech and finding phrases of interest such as abbreviations, country names, organization names and other specific entities.
- Transformation of data into a structured representation of the source content. The system distills key concepts contained in large document collections and analyzes relationships between isolated terms or phrases and documents. The transformed representation of the source text structures it for use in data exploration, clustering and predictive modeling.
- An interactive results browser that enables analysts to interactively explore concepts and relationships between documents and to make modifications to further tailor analyses.
- Full integration with SAS Enterprise Miner's data analytics software. The analytics provide comprehensive analytic tools for text and related structured data.

iv. SAS® Text Miner under the covers

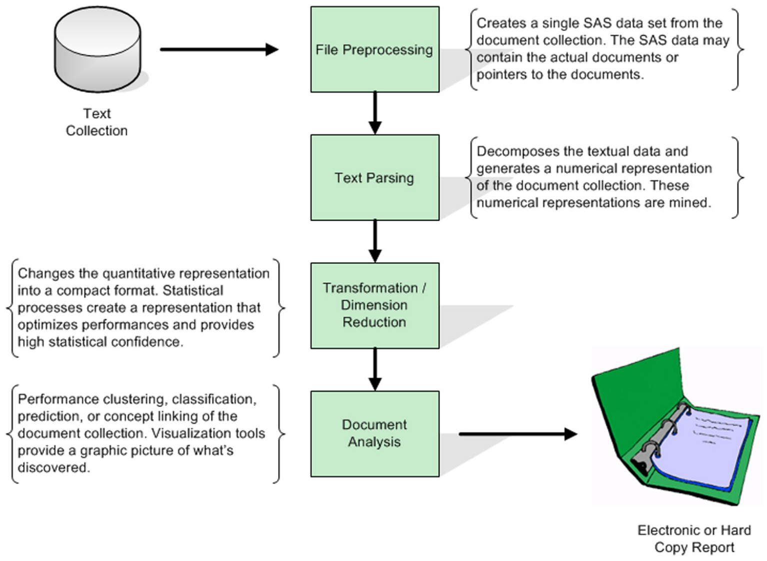
SAS Text Miner runs on most operating systems, including Linux and Windows. With the latest release, it runs on Solaris and AIX UNIX environments as well. SAS Text Miner is computationally intensive and requires dedicated hardware for optimum throughput.

SAS Text Miner relies primarily upon pattern recognition technology instead of a linguistics-centric or dictionary-based approach. SAS' text mining software produces a numerical representation of a document. In general, SAS Text Miner delivers higher accuracy when a large number of documents are processed.

SAS Text Miner is best suited to applications where hundreds or thousands of megabytes of unstructured content must be processed rapidly. The system can generate reports that make it easy for an analyst or end user to grasp the key facts, ideas and trends in the processed content.

The processes of text mining use the numerical representation to provide insights into one or more documents in a collection.

The system performs text mining by moving each document through four processing steps. Documents are acquired and preprocessed so that a single data set is formed. Next, leveraging Inxight's LinguistX Platform and ThingFinder, SAS Text Miner performs parsing and markup of the documents in the set. The processed documents are then transformed, objects tagged, and reduced to a structured form. The final stage in the process is the manipulation of the newly generated data set by analytic routines. There is no limit to the number of documents in a collection. Analysis can be performed on a single collection or across multiple collections as well as combinations of structured data from other applications and the outputs of the text mining process.


Text Collection
File Preprocessing
Creates a single SAS data set from the document collection. The SAS data may contain the actual documents or pointers to the documents.
Text Parsing
Decomposes the textual data and generates a numerical representation of the document collection. These numerical representations are mined.
Transformation / Dimension Reduction
Changes the quantitative representation into a compact format. Statistical processes create a representation that optimizes performances and provides high statistical confidence.
Document Analysis
Performance clustering, classification, prediction, or concept linking of the document collection. Visualization tools provide a graphic picture of what's discovered.
Electronic or Hard Copy Report

The SAS Text Miner performs preprocessing, text parsing, transformation, document reduction (creating a structured representation of the objects and associated metadata), and analysis.

The process flow for mining text consists of a series of actions that are performed on a static collection of data. The SAS Text Miner approach assumes that the collection or corpus to be analyzed is not updated during the parsing and other text mining processes. SAS Text Miner works through a sequence of activities in order to generate numerical snapshots of documents, clusters of objects such as major concepts in the collection itself.

There are several core processes that result in a cohesive set of data and reports about the documents processed. These basic processes are:

- Parsing: separating words and phrases in the document.
- Stemming: locating roots.

- Synonym identification: mapping roots in the document to synonyms.
- Stop list analysis: elimination of stop words via the file sashelp.stoplst.
- Term by document matrix generation: mathematical representation of each document.
- Singular value decomposition: a process to enable fast processing of numerical representations of documents.
- Clustering: a process that groups objects that are in some way related.

The results of these processes can be used by other SAS analytic routines for additional analyses and graphic expressions of the relationships identified by the algorithmic processes.

v. Reducing the manual work of text mining

Raw text is inputted to SAS Text Miner. Two pre-analysis steps are required. These are tasks usually performed by an analyst. With each release of SAS Text Miner, SAS has included routines that can reduce some of the necessary manual work. These two lists are important because the accuracy of the clustering process ultimately depends on the "knowledge" embodied in each list.

SAS Text Miner performs stemming in English, French, German and Spanish. Stemming allows variants to be grouped in one term set. An example appears in the table below.

Stem	Terms
aller (French)	vais, vas, va, allez, allons, vont
reach	reaches, reaching, reached
big	bigger, biggest
wagon	wagons

SAS Text Miner for SAS®9 is available in 15 languages. Customers receive SAS Text Miner for English and a choice of one additional language, with each customer's native language recommended.

The analyst can prepare a list of words and phrases for the system to identify. Use of a start list can reduce the amount of time required to process a corpus or repository of text.

In any collection of documents, certain terms must be identified so that SAS Text Miner processes ignore these words when processing the corpus.

Each corpus contains words that may be used as equivalents. Examples include a product name, its model number or an acronym. For example, F-150 may be a synonym for the bound phrase "light truck." SAS Text Miner uses synonyms to group concepts into a meaningful cluster.

Equivalent terms are illustrated in the table below:

Term	Parent	Category
appearing	appear	Verb
EM	SAS Enterprise Miner	Product
Employees	Employee	Noun
Administrative Assistant	Employee	Noun

SAS has developed a number of tools to assist an analyst with discovering and extracting information from a wide variety of text documents. SAS Text Miner can identify themes and concepts that are contained in the collection. The output of a SAS Text Miner analysis allows an analyst to understand the information contained in a collection of documents. An analyst or editor does not have to preprocess or pretag the content processed by SAS Text Miner.

The SAS Text Miner module generates reports. Unlike other SAS nodes, SAS Text Miner does not produce code that can be used outside the mining activity, nor does SAS Text Miner produce PMML. Nevertheless, the tight integration of the SAS Text Miner functionality in the SAS interface is attractive. An analyst familiar with SAS' data mining software can be up and running with SAS Text Miner often in as little as one day.

b. Highlighting SAS® Enterprise Miner™

i. What is the difference between SAS Text Miner and SAS Enterprise Miner?

SAS Enterprise Miner provides the process flow diagram to manage the analytical investigation as the data analysis project is pursued. The text mining application is turned on by clicking the text miner NODE located on the workspace diagram of SAS Enterprise Miner.

SAS Enterprise Miner streamlines the data mining process. The software provides an integrated system to the analyst. From a single interface, the analyst can access data, test models and generate reports for end users. SAS Enterprise Miner supports collaborative work processes.

SAS provides the most complete data mining solution available. The system supports integration with third-party enterprise applications and desktop software. Delivered as a distributed client/server system, SAS Enterprise Miner scales and is well-suited for data mining in large organizations.

SAS Enterprise Miner is designed for data miners, marketing analysts, database marketers, risk analysts, fraud investigators, engineers and scientists.

Through data mining, which SAS defines as "the process of selecting, exploring and modeling massive amounts of data," an analyst can uncover previously unknown patterns and help improve decision making across the enterprise. You may find more information at **http://www/sas.com/technologies/analytics/datamining**.

Many data mining solutions are standalone applications and often do not integrate easily into existing applications. Furthermore, scalability is a key issue, particularly when data sets can run into terabytes. Consequently, quantitative specialists must spend time accessing, preparing and manipulating disparate data. SAS Enterprise Miner automates these data preparation tasks. SAS Enterprise Miner allows analysts to model data and apply their expertise to solve specific business problems. It also ensures that the results can be deployed into scoring engines for actual implementation.

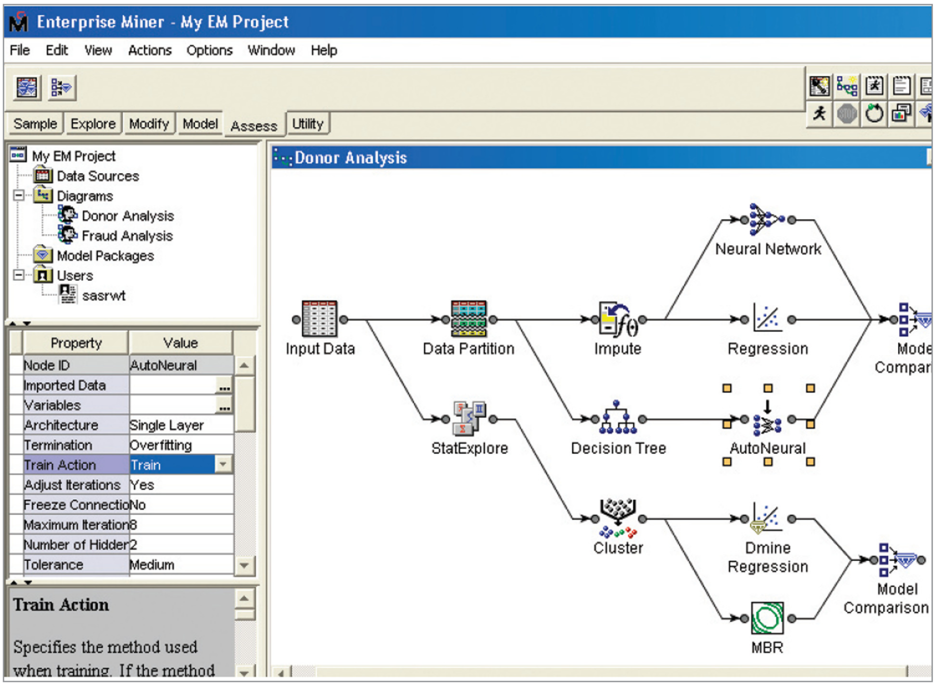
SAS Enterprise Miner includes a broad set of tools that support the complete data mining process. The functions are exposed in a graphical user interface. Most aspects of model building can be completed by pointing-and-clicking on a SAS Enterprise Miner option. SAS Enterprise Miner includes a process flow diagram system that eliminates the need for most manual coding and helps reduce the time required to construct models. SAS Enterprise Miner can generate diagrams to save the various types of analytical pattern-detecting algorithms explored. These diagrams serve as self-documenting templates. These templates can be updated as necessary and be applied to new problems without starting the analysis from scratch. Diagrams can also be shared with other analysts throughout the enterprise.

SAS Enterprise Miner offers a range of integrated assessment features. These features allow an analyst to compare the results generated by different modeling techniques. Model results can be easily shared throughout. The models reside in an SAS data mining model repository. SAS was one of the first business intelligence vendors to implement model management in an enterprise intelligence framework.

Scoring is the process of applying a model to data. The output of a data mining process consists of raw scores. SAS Enterprise Miner automates the model scoring process and supplies complete scoring code for each stage of model development. The scoring code can be deployed in numerous real-time or batch environments within SAS, on the Web or directly in relational databases.

SAS Enterprise Miner is delivered as a distributed client/server system. The subsystem eliminates the need for an organization to have niche data mining applications.

An analyst can drag and drop icons onto the process flow diagram using SAS Enterprise Miner's graphical process flow tools.



Property	Value
Node ID	AutoNeural
Imported Data	...
Variables	...
Architecture	Single Layer
Termination	Overfitting
Train Action	Train
Adjust Iterations	Yes
Freeze ConnectioNo	
Maximum Iteration8	
Number of Hidder2	
Tolerance	Medium

Train Action

Specifies the method used when training. If the method

The SAS Interface allows most functions related to programming and data manipulation to be handled within a multi-paned interface. The SAS system generates needed code automatically reducing the time an analyst spends creating scripts so that more "what if" and alternative model investigations can be conducted.

SAS Enterprise Miner implements the SEMMA data mining process. SEMMA is an acronym for SAS' core approach to data analysis: sample, explore, modify, model and assess. The idea is that an analyst can experiment with different numerical recipes, see results, make changes and then generate results from the optimal model often described by SAS as the champion model. SEMMA is a logical organization of the functions in the SAS Enterprise Miner subsystem. SEMMA provides a flexible framework for conducting the core tasks of data mining.

SAS Enterprise Miner uses a Java client and the SAS server architecture. The result is that the mining computational server is separate from the user interface. An analyst can configure an installation that scales from a single-user system to very large enterprise solutions with no changes to the model or recoding. Certain process-intensive server tasks such as data sorting, summarization, variable selection and regression modeling are multithreaded. SAS Enterprise Miner can automatically distribute the workload over multiple CPUs. Its processes can be run in parallel and asynchronously. Large or repetitive training or scoring processes can be scheduled for batch processing during off-peak demand hours on the analytical server without programming.

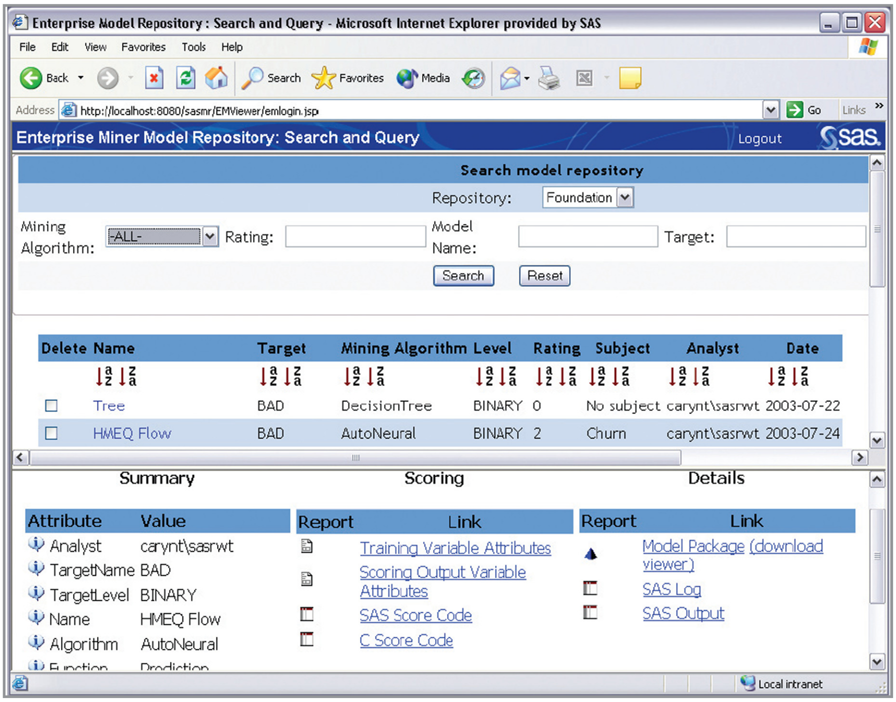
SAS Enterprise Miner bakes in advanced predictive and descriptive modeling tools and algorithms. An analyst can implement decision trees, neural networks, autoneural networks, memory-based reasoning, linear and logistic regression, clustering, associations, and time series among others, by pointing and clicking on a menu choice or a radio button. With multiple models and algorithms ready to run in a single subsystem, SAS Enterprise Miner allows an analyst to explore data and the outputs of different models from a visual user interface.

SAS Enterprise Miner includes assessment capabilities. These help ensure that a common framework for comparing different modeling techniques is available to the analyst.

SAS Enterprise Miner includes tools to help the analyst perform such tasks as sampling, data partitioning, missing value imputation, clustering, merging data sources, dropping variables, customized SAS language processing via the SAS code node, variable transformation and filtering outliers. Extensive descriptive summarization features are included, as well as advanced visualization tools that permit examination of large amounts of data in multidimensional plots and graphically compare modeling results.

The platform-independent Java-based GUI of SAS Enterprise Miner provides enhanced statistical graphics with flexible actions and controls. A Java graphics wizard is also provided for developing customized graphs. Plots and underlying tables are dynamically linked by SAS Enterprise Miner to allow interactions with the model and other components of the system.

The SAS Enterprise Miner environment features integrated assessment functions. These allow an analyst to compare the results of different modeling techniques. An analyst can gauge a model's effectiveness in statistical terms and in more concrete measures such as money saved or new sales revenue.


Enterprise Model Repository : Search and Query - Microsoft Internet Explorer provided by SAS
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address
Go
Links
Enterprise Miner Model Repository: Search and Query
Logout
sas.
Search
Reset
Local intranet

Address: http://localhost:8080/sasmr/EMViewer/emlogin.jsp

Search model repository

Repository: Foundation

Mining Algorithm: -ALL- Rating: Model Name: Target:

Delete	Name	Target	Mining Algorithm	Level	Rating	Subject	Analyst	Date
□	Tree	BAD	DecisionTree	BINARY	0	No subject	carynt\sasrwt	2003-07-22
□	HMEQ Flow	BAD	AutoNeural	BINARY	2	Churn	carynt\sasrwt	2003-07-24

Summary

Attribute	Value
Analyst	carynt\sasrwt
TargetName	BAD
TargetLevel	BINARY
Name	HMEQ Flow
Algorithm	AutoNeural
Function	Prediction

Scoring

Report	Link
	Training Variable Attributes
	Scoring Output Variable Attributes
	SAS Score Code
	C Score Code

Details

Report	Link
	Model Package (download viewer)
	SAS Log
	SAS Output

Scoring data often appear in a series of tables. SAS's approach is to make the scoring data for various runs accessible in a point-and-click interface.

SAS Enterprise Miner features a form of version control for models. Analysts can keep track of the models that have been updated and compare the improvements of accuracy over time. Models developed for other projects can be located and their components reused in a new project.

SAS Enterprise Miner allows the analyst to save schematics and flow diagrams. It exports these objects as XML files. An analyst can create a compressed model results packages containing the data used in a data mining process flow, including data preprocessing, modeling logic, model results and scoring code.

These model result packages can be registered to the SAS Metadata Server for subsequent retrieval and querying by data miners, business managers and data managers via a Web-based model repository viewer. When introduced in 2002, SAS had the only Web-based system for effectively managing and distributing large model portfolios throughout the organization.

ii. Model scoring

In some organizations, the analyst makes the model available to managers in operating units. Scoring – sometimes called weighting of different factors in a model – requires considerable expertise. Scoring in some text mining systems is a time-consuming process. In some cases, the analyst will have to write or edit scoring scripts. The SAS Enterprise Miner environment automates virtually all of the scoring code process. If a manual adjustment is required, SAS provides scoring code in SAS script, C, Java and PMML (predictive model markup language). The SAS implementation captures the code of an analyst's models, supports the import and export of PMML models for Naive Bayes and association rules models, and codes for preprocessing activities such as data normalization.

Once the scoring code has been captured, an analyst can:

- Score the production data set within SAS Enterprise Miner.
- Export the scoring code and score on a different machine.
- Deploy the scoring formula in batch or real-time on the Web or directly in relational databases.

iii. Customizing applications

SAS Enterprise Miner can be customized and extended. The default tool library of SAS Enterprise Miner allows an analyst to script additional functions using SAS code and XML logic. SAS supports a Java API that can be embedded into customized applications. SAS makes it possible for an analyst to build a proprietary analytical application such as having an OLAP report capability linked with a third-party application within the same interface.

c. How do these products interrelate?

The flagship product SAS Text Miner features a comprehensive range of text mining tools to speed search and data analysis, including:

- Noun group extraction support for all languages.
- Additional entity types.
- An enhanced TMFILTER macro tool.
- UNIX support.
- More robust synonym list processing and parsing, including a revised DOCPARSE procedure and DOCSCORE DATA step scoring function.

5. SAS® Text Miner Feature Snapshot*

Feature	Comment
Analytics functions	A comprehensive range of applications in a diverse range of industries exists. The most popular include (1) fraud detection, (2) warranty early warning and (3) supplier relationship management statistics. Comprehensive text preprocessing capabilities include: • Capture and distill the most important underlying information within a document collection. • Default or customized stop lists for each language to remove terms with little or no informational value. • Automated spelling correction. • Stemming to identify root words. • Part-of-speech tagging based on sentence context. • Noun group extraction for identifying phrase-level concepts such as "competitive intelligence." • User-defined multiword tokens, such as "point and click." • User-customized and default synonym lists. Compound word splitting into distinct subterms.
API	APIs are provided with complete documentation and sample code.
Controlled Vocabulary Support	The system can make use of existing dictionaries as well as develop one from scratch as documents are analyzed and common terms are surfaced. Instead of selling industry-specific dictionaries, SAS offers special routines that allow domain experts to rapidly tailor their own stop/start lists of synonyms with user-friendly interactive point-and-click programming interfaces. For a predictive modeling exercise, significant effort with industry-specific dictionaries can degrade results. If a customer has his or her own dictionary, SAS can incorporate that into the model.
Direct third-party support	Third-party applications can make use of SAS outputs. The APIs are used to link SAS Text Miner with other systems and software. Score code can be saved as a stored process and deployed through a familiar BI client such as Excel, SAS Web Report Studio, SAS Information Delivery Portal and SAS Enterprise Guide.
Document conversion support	SAS Text Miner can read text stored in a variety of document formats, such as PDF, ASCII, HTML, Microsoft Word and WordPerfect.
Entity extraction	People, places, events, dates and other objects can be extracted.
Language support	Danish, Dutch, English, Finnish, French, German, Italian, Japanese, Korean, Norwegian (Bokmal), Portuguese, Spanish, Swedish, Traditional Chinese and Simplified Chinese. • Support for Latin-1, Double Byte Character and UTF-8 encodings. • European languages (Latin-1 encoding): Danish, Dutch, English, Finnish, French, German, Italian, Norwegian (Bokmal), Portuguese, Spanish and Swedish. • Far-Eastern languages (Double Byte Character Support): Japanese, Korean, Simplified Chinese and Traditional Chinese. Encoding support for Unicode UTF-8 allows analysis of non-English interfaces and languages that listed above

Professional services	SAS offers free Web tutorials, on-site education and training, on-demand Web seminars, engineering, conferences, academic initiatives for schools and universities, and consulting services.
Programming languages	Major programming languages are supported. Graphical interfaces are available to modify certain system functions.
Ready-to-run modules	SAS systems require installation and some setup prior to operation. Minimum processor speed is 1 GHZ. Memory requirements: 512 MB server. Disk space required: 40 MB thin Java client, 300 MB SAS client. 300 MB server – average Win XP install.
Platform	SAS runs on Windows (32-bit) AIX and Solaris (64-bit).
Crawler	The system includes tools to acquire information. Additionally, the system can process information pushed to watched folders.
Standards support	The SAS Text Miner system supports Java and XML standards. SAS is actively participating in the UIMA standards discussions with IBM and other alliance partners, all of whom join in the PMML discussions at DMG.org and the Grid Computing Forum. SAS is monitoring development efforts to identify and select meaningful metadata so that when it does integrate with a UIMA-compliant standard of application, it will be useful to SAS customers, especially for analysis purposes.
Taxonomy management tools	The system provides a graphical interface to manage classification systems and represent the class/subclass relationships in a hierarchical fashion displaying documents in a “tree.”
Taxonomy support	SAS allows the taxonomy be “discovered” with an unsupervised method, hierarchical clustering. We also “discover” the labels for the classes and subclasses. Or SAS Text Miner can follow a predefined classification system, if available.
Third-party software support	Most third-party and proprietary systems can be supported by SAS Text Miner and SAS Enterprise Miner.
Visualization functions	SAS provides a range of data visualization tools. The concept-linking diagrams have been tailored specifically for SAS Text Miner. User-directed concept linking provides flexible navigation to visualize complex hidden relationships between terms, phrases and entities (such as people and place names). Concept linking, now integrated with the Interactive Results Browser for enhanced usability and greater insight, helps detect patterns in a clear visual fashion that may not have otherwise been observed.

*Features described reflect the November 2006 release of SAS Text Miner.

6. Conclusion

SAS has done the work required to ease the integration of text with structured data elements through drag-and-drop or point-and-click functions. SAS Text Miner provides a full range of predictive modeling tools that can unearth opportunities for timely exploitation. Benefits include

- Data in a SAS System can be traced. The notion of data lineage is important when the organization must operate within the guidelines of Sarbanes-Oxley, Basel III, and other compliance and governance requirements.
- A large organization can develop a comprehensive, customized text mining system without the cost and time required to integrate third-party tools into a cohesive system.
- SAS has continued to add functionality to the SAS Text Miner node; for example, SAS has added a credit scoring function that can use a wide range of numeric and textual information.

SAS Text Miner's approach sets it apart from many business intelligence and text mining systems. Because SAS Text Miner is a component residing within a larger data management and analytic environment, an organization using SAS Text Miner can repurpose models, streamline "what if" and alternative model analysis, and merge outputs from SAS Text Miner with data from numeric or hybrid systems. These built-in services slash the time and programming required to achieve similar goals using software from multiple vendors.

The programmer analyst with a strong foundation in statistics will revel in SAS' technology and its power. A person who invests the time to learn the SAS approach will find that SAS delivers usable, flexible and scalable analytic tools.

End users benefit from visual presentations of complex data and point-and-click interfaces created with SAS' built-in tools. Adequate computational horsepower, system specialists, and analysts with the requisite programming and statistical expertise are all that's needed to make SAS hum.

In summary, integrating text-based information with structured data enriches your predictive modeling capabilities and provides new stores of insightful information for driving your business and research initiatives forward. Text mining supports a wide variety of applications, such as categorizing huge collections of call center data, Web text and blogs analysis, finding patterns in customer feedback or employee surveys, detecting emerging product issues, analyzing competitive intelligence reports or patent databases, and classifying reports and company information by topic or business issue.

Achieving industry-leading status almost upon its introduction, these SAS data mining technologies continue to receive rave reviews from industry experts and users alike. Keep your eyes on SAS.

For more information on SAS Text Miner, visit:

www.sas.com/technologies/analytics/datamining/textminer

7. Appendix

Birds-eye view	
Product	SAS Text Miner (requires SAS Enterprise Miner and SAS/STAT® licenses)
Platform	Text mining toolkit – runs on Microsoft Windows, UNIX.
License Fee	Text Mining add-on starts at about $30,000.
Use Cases	Transforms textual data into a usable, intelligible format that facilitates classifying documents, finding explicit relationships or associations between documents, clustering documents into categories and incorporating text with other structured data to enrich predictive modeling endeavors.
Key Features	A variety of advanced analytics – statistical, heuristics, neural nets, pattern decisions and predictive forecasting data mining technologies.
Caveats	This SAS technology runs on a client/server platform. Installation must be done carefully to define the proper metadata architecture.
Similar to	Like products from SPSS, TEMIS and IBM, the SAS Text Miner product performs textual analytics. Unlike competitors, SAS excels in the integration of the structure and unstructured data for predictive scoring models.
Points to Note	SAS provides an industry-standard, integrated statistical operating system and the SAS Text Miner module to “bridge the gap between data and text.”