

Paper 102-2013**A Tale of Two SAS Technologies – Generating Maps of Topical Coverage and Linkages in SAS User Conference Papers****Denise Bedford, Kent State University, OH****Barry de Ville, SAS Institute, Cary, NC****Rich La Valley, Strategic Technology Solutions, Herndon, VA****ABSTRACT**

This paper discusses how two SAS technologies – Text Analytics and Content Categorization Suite -- were used to generate comprehensive and dynamic categories and clusters of the entire corpus of SAS user presentations from inception to the present. The goal of the two approaches is to improve access to the conference proceedings for SAS users and conference attendees in particular. The research focuses on the Conference Section access point and considers how text analytics and content categorization might enrich end user access.

The findings of this research suggest that both Text Miner and the SAS Content Categorization Suite are powerful tools that can be used independently or in complement to improve access. Similarities were observed for Technology Solutions. This was not a surprising result given the dense treatment of technology issues at SAS Conferences and the technical expertise of conference attendees and Global Forum members. The Industry perspective was much less obvious and the treatment of industry concepts was much sparser in both the text analytic generated clusters and the knowledge engineered categories. The research suggests that there are clear advantages and tradeoffs in working with both technologies.

INTRODUCTION

For the past 37 years, SAS and SAS users have shared their work and knowledge at an annual conference. The papers given at each conference have been captured in annual proceedings. Until recently, the majority of these papers were only available as print publications handed out to users who attended the conference. This is a common conference practice. A second access point was a website created by Lex Jansen for the SAS user community (Jansen 2003). In 2011, the SAS Global Users Group launched and completed a project to digitize, scan and index the historical collection of conference papers. This project resulted in a digital repository of 9,134 papers (La Valley, Jansen, Lafler 2013). This repository was the focus of this research and constituted the data set.

Having a digitized and semantically readable collection of papers made it possible to expand access to the papers. The existing SUGI/SAS Global Forum Conference Papers search is a multifaceted search capability which covers: Conference Section, Company, Country, Paper Type, Skill Level, and Speaker. Each facet has a different structure, a different knowledge base, and a distinct behavior. Conference Section, the focus of this research, represents the section to which the paper was assigned by the conference program committee. While this is an important access point, it presents limitations. Conference Sections change over time, with new sections appearing, others disappearing, and some changing focus. Papers are rarely one-dimensional. While a paper may be assigned to a section to support conference scheduling, a section may not represent all of the concepts or solutions that are discussed by the authors. While author assigned keywords may compensate for this limitation, they present a degree of scatter across concepts. Author assigned keywords reflect the authors perspective of the work, and may fail to surface aspects which are important to others.

For these reasons, the research team undertook an exploratory investigation to learn how access to the SUGI/SAS Global Forum Conference Papers might be enhanced. Two methods were considered. Method 1 focused on SAS Text Analytics technologies. Method 2 focused on the SAS Content Categorization Suite. The sections below describe the two methodologies and their results.

BACKGROUND

As we have seen, in text analytics we can attempt to exploit our knowledge of the domain coupled with our ability to determine the kinds of concepts, characteristics and categories we are looking for and to refine pattern-matching tools such as taxonomies. Because our domain knowledge provides the leverage that we need to define what we are looking for, we can call this approach “domain-driven”. As we have seen, one of the foundations that enable us to deploy our domain knowledge in this fashion consists of various linguistics engines that can consume textual data and can recognize parts of speech, particular kinds of entities and so on.

Discovery and learning based approaches to text analytics, like the domain-driven approaches described above, also rely on a foundation layer of linguistics engines and also use linguistic products such as domain-driven taxonomy definitions. The distinction between discovery-driven text products and domain-driven text products lies in relaxing assumptions and fore-knowledge about how various text products may operate together to produce semantic meaning. Discovery relies heavily on various search and pattern-recognition techniques that are designed to discover and highlight associations and relationships between various elements of text. In discovery, we attempt to extract meaningful, usual novel, and often unanticipated relationships and composites that can inform the analysis and enrich our understanding of the text corpus.

In past SUGI-SGF conferences we have seen many discovery-based approaches to text analytics: (Cerrito, 2004, DeVille, 2006, Allbright et al. 2007). Last year, we saw a forerunner to the current topic in a presentation that demonstrates the application of discovery-based text mining to tracking trends in the SUGI-SGF papers themselves (Shaik, et. al., 2012). A prominent introduction to text mining features the text analysis of the SAS user groups' papers as an illustrative use case (SAS Institute, 2004).

Semantic analysis using the SAS Content Categorization Suite was first discussed at a SUGI-SGF conference by Bedford and LaValley (2011). Whereas text analytics applies technologies to text for the initial discovery, the SAS Content Categorization Suite leverages human knowledge and knowledge structures to guide discovery (*Bedford, October 2012*), (*Bedford ASIR 2012*), (*Bedford June 2012*), (*Bedford and Gray May 2012*), (*Bedford and Gracy October 2012*), and (*Bedford and Gardner, 2011*).

GENERAL RESEARCH METHODOLOGY

As the paper title suggests, two technologies were used to support two approaches to semantic analysis. Method 1 leveraged text analytics and statistical categorization to discover categories. Text Analytics is defined as the process of analyzing unstructured text, extracting information, and transforming information into structured or semi-structured information and the discovery of patterns that lead to semantic understanding of what is in the unstructured text. In the past 37 years, SAS has developed and provided numerous tools to analyze and discover information in structured data that is stored in databases or other data stores.

Method 2 leveraged the SAS Content Categorization Suite and focused on knowledge engineering and deep semantics. The Categorization Suite includes five semantic methods, including rule based concept extraction, grammar based concept extraction, rule based categorization, statistical categorization, and automated summarization. For the purpose of this research, two of the methods were used: grammar-based concept extraction and rule-based categorization.

Each technology was applied to the digitized collection of papers. Some steps in the processes were similar, while others were unique. Table 1 identifies common and divergent steps in the processes followed by two of the authors.

Table 1. Comparison of Methodologies

Steps	Method 1 (aka "discovery")	Method 2 (aka "engineering")	Common and Divergent Steps
1	Assemble Document Corpus	Assemble Document Corpus	Same Process
2		Knowledge Engineering of High Level Categories	Unique Step
3	Select POS units of analysis	Define Grammatical Concept Extraction Methods	May be similar if both technologies have NLP foundation
4	Define Clustering/Association Rules and Parameters	Construct Rule-Based Categorization Profiles	Different processes that require different technologies
5	Run Text Analytics Software and Generate Statistical Topics/Clusters	Apply Semantic Profiles to Documents in the Corpus	Variant processes – one generates categories as the result; the other generates explicit exportable metadata for

Steps	Method 1 (aka “discovery”)	Method 2 (aka “engineering”)	Common and Divergent Steps
			each document.
6a	Validate Clusters Through Human Review	Review Categorization of Documents	One process validates the categories, the other process validates the assignment of documents to categories
6b	Adjust the Topic/Clustering Rules and Parameters	Adjust the Categorization Profiles at Concept Level	Same Process – both methods include a validation and refinement step
7		Review Deep Concept Indexing of Documents	This step is not covered in Method 1
8		Export Concepts and Categories to Navigation and Search Applications	This step is not covered in Method 1

The comparison of methods surfaces varying levels of effort and knowledge required. The SAS Content Categorization Suite requires knowledge engineering and direction from the researcher. The SAS Text Analytics method relies largely on undirected discovery methods. Each of these methods is discussed in detail in the sections below.

METHOD 1: DISCOVERY AND LEARNING PROCESS OVERVIEW

In the discovery and learning approach presented here we begin – as all discovery approaches do -- by consuming a baseline data stream of standard linguistic products. Each document in our collection is run through a linguistic pre-processor so that parts of speech and important entities in the SUGI-SGF proceedings data set are identified for each document in each year of this analysis. In this respect, the discovery approach begins the same way a domain-driven approach begins. Initially, the approaches are equivalent; in fact, any and all domain-driven text products can be moved into the discovery stage of analysis for a further discovery step. As we will see later, discovery products can be passed back into domain-driven approaches as a way of improving these approaches and as a way of continuously refining and enhancing our text analytic capacities.

Our main goal in this example of the discovery process is to present what we call a “semantic map” that characterizes the content of the proceedings corpus during the years of our analysis. This semantic map is driven exclusively by the operation of discovery algorithms: we make no assumptions about which terms or documents occur together. We hope to demonstrate that this approach will show clusters of semantically-related components which, when examined, provide an informative description of the various kinds of presentations that we have seen over the years.

As a next step we will demonstrate the discovery of text topics to show how documents can be characterized with reference to the topics of conversation that they contain. So, in addition to providing a map of the general types of presentations we will look at the sub-document level of text to extract combinations and associations of terms which tend to be presented together *within* the presentations that occur over the years.

Once we have done this we will use the topics that we have discovered to update the general semantic map that we began with so as to more clearly describe what the map shows us, based on an overlay of the topics that we have discovered.

At this point, we hope to have demonstrated that, in a relatively automatic way, we have learned something novel about the collection of documents that we might not have anticipated on the basis of domain knowledge alone.

As a final step, having discovered new meaning, we show how the discovered text products can be converted into linguistic rules and how these linguistic rules can be used to enrich the front end document scanning that is done in the future. As shown in Figure 1, this demonstrates the closed-loop, synergistic nature of domain-driven and discovery-driven approaches and points the way to a more effective overall process that is one of the unique hall marks of the SAS text analytics solutions.

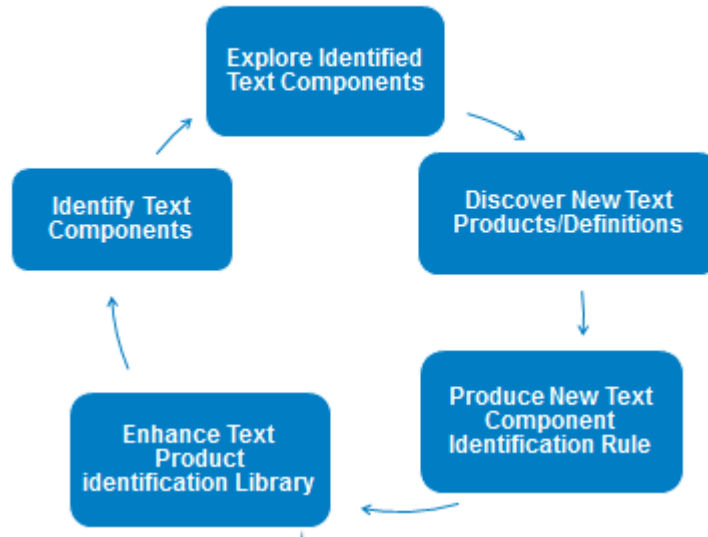


Figure 1: The "Virtuous Cycle" Of Domain-driven and Discovery-driven Text Analytics

This last step of "closing the loop" is often referred to as a virtuous circle or virtuous cycle: "A condition in which a favorable circumstance or result gives rise to another that subsequently supports the first." as referenced in Wikipedia.

DETAILED PROCESS DESCRIPTION

DATA PREPARATION

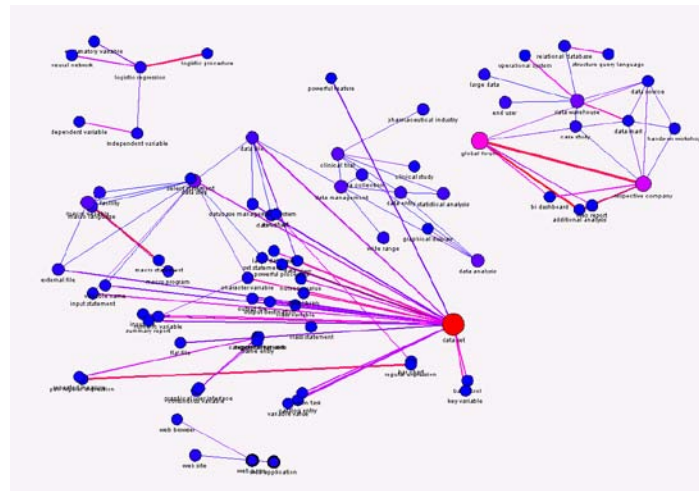
As described in an earlier section, we used %tmfilter to pull in text from all .pdfs in year range 1989 – 2012. The distribution of years is shown below. As we can see the number of papers is relatively constant over these years (although the number of papers appears to have been rising in the last few years). Table 2 summarizes the counts and the percentages in each year

Table 2. Year Counts

Year	Frequency Count	Percent		Year	Frequency Count	Percent
1989	321	4.44%		2001	262	3.62%
1990	286	3.96%		2002	278	3.85%
1991	303	4.19%		2003	278	3.85%
1992	268	3.71%		2004	265	3.67%
1993	265	3.67%		2005	258	3.57%
1994	274	3.79%		2006	261	3.61%
1995	264	3.65%		2007	375	5.19%
1996	275	3.80%		2008	384	5.31%
1997	261	3.61%		2009	406	5.62%
1998	253	3.50%		2010	323	4.47%
1999	308	4.26%		2011	378	5.23%
2000	271	3.75%		2012	411	5.69%
				Total	7228	100.00%

The text mining “Parse” node in Enterprise Miner allows us to identify various parts of speech (Noun, verb and so on) as well as specific entities (Person, Location, for example). For purpose of the illustration here we elected to significantly restrict the text parse products that were brought into Text Miner for analysis. Since, for purpose of illustration, we are most interested in producing a classification of the various kinds of papers that are presented we elected to focus on parts of speech that can serve as subjects of a sentence, typically a noun or noun phrase. In the interest of economy of presentation we have selected only noun phrases. Our intent is to derive a high level summary of significant subjects to serve as a classification of papers presented in the selected time period of 1989 – 2012.

From these noun phrases we used the Enterprise Miner “Associations” node to produce a network diagram which will serve as a “semantic map” to describe the overall content of SGF-SUGI proceedings during our target 1989 – 2012 time period.



CONSTRUCTION OF TEXT TOPICS

The approach employed by SAS Text Miner leverages the capabilities that have been built into Text Miner since its initial release, notably the construction of singular value decompositions (SVDs) that produce a structure-preserving summary of higher dimensional document representations in lower dimensional space. These SVD composite can be treated as factors or principal components and so can be manipulated through the construction of factor, or aggregate, scores which can then be clustered, for example.

5

of the composite expression line up on the rotated factor dimension in the most interpretable and meaningful way. This provides us with a substantial flexibility in constructing the topics or concepts that we extract.

The Text topics node has the further capability of:

- Determining whether we retrieve single term or multi-term composites
- Determining the n-most significant composites to retrieve

We have the capability to manually-edit the weight of a given component to a composite (including assigning a zero weight to have the component excluded).

Later, by virtue of a **Text Rule Builder** capability we will see that we have the ability to render Topics – and the associated mathematical expressions – in Boolean form that can be directly employed in the front end linguistic engineering workstation. As we will see, the **Text Rule Builder** creates classification rules in native Content Categorization code.

In forming the topics for this demonstration we used Noun Groups, as shown in the semantic map above. Clearly, this facilitates the comparison of the two sets of results. We have the capability of finely fitting a large number of text topics to the existing set of documents. For purpose of this demonstration we chose to extract a relatively small set of top-most topics so that the results would be more easily interpretable. The text topics that we created are as follows in Table 3:

Table 3. Text Topics

ID	Descriptors	Terms	Documents
2	+data set, +data step, +data file, +set statement, +external file	65	927
8	output delivery system, +data step, +operate system, +web page, statistical graphics	165	793
7	+clinical trial, +logistic regression, +confidence interval, +independent variable, output delivery system	214	717
6	+clinical trial, +data entry, +operate system, +data management, clinical data	171	690
1	other brand, +product name, +respective company, +service name, +global forum	27	483
4	+macro variable, +macro language, +macro facility, macro processor, +macro program	93	351
3	+data warehouse, +data mart, data integration, +operational system, data quality	93	273
5	enterprise guide, +hands-on workshop, +workshop participant, hands-on experience, +web report	84	162

The topics are displayed from most common (most documents covered) to least common.

From top to bottom we can eyeball the documents, arbitrarily assign a textual label (descriptor) that appeals to us and create a global content map as follows in Table 4

Table 4. Textual Labels

Topic and Descriptor
Topic 2 – Data Step
Topic 8 – Output Delivery System (ODS)
Topic 7 – Clinical Trials – Statistics, Analysis, Reports
Topic 6 – Clinical Trials – Data Assembly, Data Management
Topic 1 – Global Forum Copyrights/Endnotes
Topic 4 – Macros – Macro Language and Utilization
Topic 3 – Data Warehouse, Data Quality
Topic 5 – Enterprise Guide – Workshops, Web Reports

A TOPIC-BASED SEMANTIC MAP

We take the topics that we have described above and place them on the same set of documents that we used to derive the initial semantic map, shown above. This enables us to produce a semantic map with topic references, shown below.

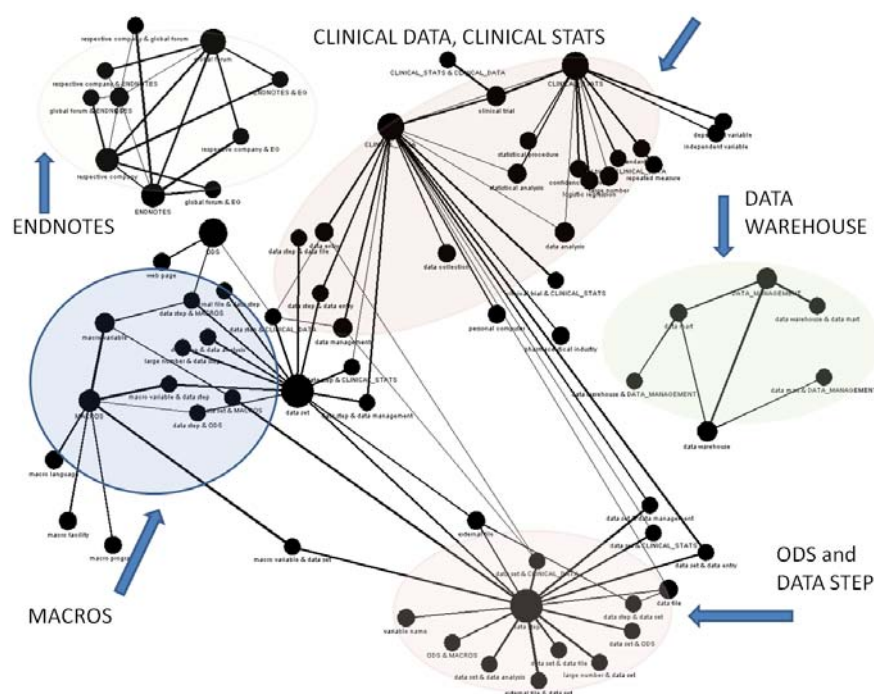


Figure 3: Semantic Clusters Co-located with Major Topic Descriptors

To make it easier to identify the topics they are presented in CAPITAL letters. The figure has been annotated by colored regions to accompany each of the five topics that are identified here. In the interest of the production of an economical display, only the top associations are presented in this diagram. This presentation results in the identification of an overlay between 7 of the topics identified above with the underlying clusters of associated noun phrases. The “Enterprise Guide” topic is not presented in this diagram since it was not related with high-frequency noun phrase associations.

We do see that the “Macros” and “Endnotes” topics align as relatively distinct clusters on the left hand side of the diagram. We also see that “Clinical Data” and “Clinical Statistics” topics cluster together at the top of the diagram while “Data Step” and “ODS” cluster together at the bottom. The “Data Warehouse” topic cluster is distinctly shown in the right-hand side of the diagram, along with the associated noun phrases.

If we re-form our semantic map based on this high-level overview (and ignore the endnotes cluster as an artifact of the editorial description of the document) then we recover a revised high-level semantic map of the SUGI-SGF proceedings that appears as follows:

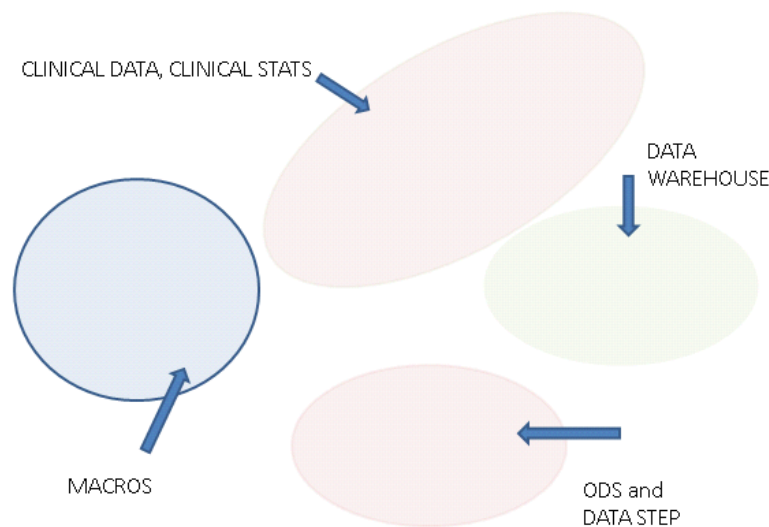


Figure 4: Content and Distribution of High Level Content 1989 –2012

LEARNING AND RULE BUILDING

Now that we have identified the important topic-noun phrase discovery products we will likely have an interest in re-casting these text products as linguistic rules that can be used in the front end parsing of any future documents. In this way we can ensure that, once a text product is “discovered”, it can be turned into a production product and can be automatically identified in future analytic scenarios. This step completes the “virtuous cycle”, defined above.

Text Miner provides a rule builder facility that enables us to produce production-ready rules to accomplish this purpose. The **Text Rule Builder** node generates an ordered set of rules from small subsets of terms that together are useful in describing and predicting a target variable. Each rule in the set is associated with a specific target category that consists of a conjunction that indicates the presence or absence of one or a small subset of terms (for example, “term1” AND “term2” AND (NOT “term3”)). A particular document matches this rule if and only if it contains at least one occurrence of term1 and of term2 but no occurrences of term3. This set of derived rules creates a model that is both descriptive and predictive. When categorizing a new document, the rule building process proceeds through the ordered set and chooses the target that is associated with the first rule that matches that document. The rules are provided in the syntax that can be used within SAS Content Categorization Studio, and can be deployed there. For more information about the **Text Rule Builder** node, see the SAS Text Miner help.

Presented here are the discovered rules for categorizing a document as containing Text Topic 2. Recall from above that Text Topic 2 is labeled “Data Step” and contains the terms “data set”, “data step”, “data file”, “set statement” and “external file”. Since the Text Rule Builder can define a topic by the terms that it contains as well as the terms that it does not contain we will see rules that contain the defining terms we see from the topic computation but also terms that we discover are not related to the topic. This increases the power and precision of the derived rules.

Note: ~ means logical not.

Table 5. Text Topic 2

F_TextTopic_2 =1 ::
(OR
, (AND, (OR, "data steps", "data step"), (OR, "data sets", "data set"))
, (AND, (OR, "data set", "data sets"), (OR, "external file", "external files"), (NOT, "maximum likelihood"), (NOT, "statistical graphics"), (NOT, (OR, "mixed procedure", "mixed procedures"))))
, (AND, (OR, "data sets", "data set"), (OR, "data file", "data files"), (NOT, "maximum likelihood"), (NOT, "statistical graphics"), (NOT, (OR, "mixed procedures", "mixed procedure"))))
, (AND, (OR, "data set", "data sets"), (OR, "larger data", "largest data", "large data"), (NOT, "maximum likelihood"), (NOT, "statistical graphics"), (NOT, (OR, "mixed procedures", "mixed procedure"))))
, (AND, (OR, "data set", "data sets"), "input data", (NOT, "maximum likelihood"), (NOT, "statistical graphics"), (NOT, (OR, "mixed procedures", "mixed procedure"))))
, (AND, "<", (OR, "data set", "data sets"), (NOT, "maximum likelihood"), (NOT, "statistical graphics"), (NOT, (OR, "mixed procedure", "mixed procedures"))))
, (AND, (OR, "data sets", "data set"), (OR, "data library", "data libraries"), (NOT, "maximum likelihood"), (NOT, (OR, "mixed procedures", "mixed procedure"))))
, (AND, (OR, "data sets", "data set"), (OR, "variables names", "variable name", "variable names"), (NOT, "maximum likelihood"), (NOT, (OR, "mixed procedures", "mixed procedure"))))
, (AND, (OR, "data sets", "data set"), (OR, "set statements", "set statement"), (NOT, "maximum likelihood"), (NOT, (OR, "mixed procedures", "mixed procedure"))))
, (AND, (OR, "data entries", "data entry"), (OR, "data sets", "data set"), (NOT, "maximum likelihood"), (NOT, (OR, "mixed procedures", "mixed procedure"))))
, (AND, (OR, "data sets", "data set"), (NOT, "maximum likelihood"), (OR, "data structures", "data structure"), (NOT, (OR, "mixed procedures", "mixed procedure"))))
, (AND, "=", (OR, "data set", "data sets"), (NOT, "maximum likelihood"))
, (AND, (OR, "data set", "data sets"))

Below we see an analysis of the generated rules found in Table 6:

Table 6. Rules Generated

True Positive/Total	Remaining Positive/Total	Rule	Estimated Precision	Sample Precision	Sample Recall
114/114	927/6,659	data set & data step	0.97	1.00	0.12
39/39	813/6,545	data set & ~maximum likelihood & ~statistical graphics & ~mixed procedure & external file	0.91	1.00	0.17
48/49	774/6,506	data set & ~maximum likelihood & ~statistical graphics & ~mixed procedure & data file	0.90	1.00	0.22
26/26	726/6,457	data set & ~maximum likelihood & ~statistical graphics & ~mixed procedure & large data	0.87	1.00	0.24
25/25	700/6,431	data set & ~maximum likelihood & ~statistical graphics & ~mixed procedure & input data	0.86	1.00	0.27
57/60	675/6,406	data set & ~maximum likelihood & ~statistical graphics & ~mixed	0.88	0.99	0.33

		procedure & <			
17/17	618/6,346	data set & ~maximum likelihood & ~mixed procedure & data library	0.82	0.99	0.35
17/17	601/6,329	data set & ~maximum likelihood & ~mixed procedure & variable name	0.82	0.99	0.37
14/14	584/6,312	data set & ~maximum likelihood & ~mixed procedure & set statement	0.79	0.99	0.39
25/26	570/6,298	data set & ~maximum likelihood & ~mixed procedure & data entry	0.83	0.99	0.41
18/19	545/6,272	data set & ~maximum likelihood & ~mixed procedure & data structure	0.79	0.99	0.43
106/126	527/6,253	data set & ~maximum likelihood & =	0.81	0.95	0.55
421/559	421/6,127	data set	0.75	0.85	1.00

The words in the Rule column have the corresponding estimated precision at implying the target, **Text Topic 2**. In the second column above, the True Positive (the first number) is the number of documents that were correctly assigned to the rule. The Total (the second number) is the total positive.

In the third column above, the Remaining Positive (the first number) is the total number of remaining documents in the category. The Total (the second number) is the total number of documents remaining.

In the above example, in the first row, 114 of the documents were assigned to the rule “data set and data step” (114 were correctly assigned). This means that if a document contains the word “data set and data step” and you assign all those documents to Text Topic 2, 114 out of 114 will be assigned correctly. We can also see that there are 927 out of 6,659 documents that contain Text Topic 2 that were not identified by this rule. So what we see is that the estimated precision of the rule is .98 and the sample precision is 1.00.

In the next row, there are 927 Topic 2 documents left that can be evaluated for rule assignment, out of a total of 6,659 – 114 = 6,545 documents. In this second row, 39 documents are correctly assigned to the rule “data set & ~maximum likelihood & ~statistical graphics & ~mixed procedure & external file” This means that if a document contains the term “data set and external file” and you assign all those documents to the Topic 2 newsgroup, 39 of 39 will be assigned correctly.

METHOD 2: KNOWLEDGE ENGINEERED CONTENT CATEGORIZATION AND DEEP SEMANTICS

Method 2 leveraged the SAS® Categorization Suite. This Suite of products supports semantic solutions, including: (1) grammatical concept extraction; (2) rule-based concept extraction; (3) dynamic (statistical) categorization; (4) rule-based categorization; and (5) rule based sentence extraction. Two of the semantic solutions were used in this research: rule-based categorization and grammatical concept extraction.

SAS® categorization suite allows the designer to define specific sets of grammatical units for extraction, using a standard tag set for natural language processing. The research leveraged this technology to identify noun-phrases from pertinent documents organized into training sets. The extracted noun-phrases were used to construct the deep rule-based categories.

RULE BASED CATEGORIZATION

Rule-based categorization involves defining a classification structure and constructing the knowledge base for each class. In this case, two categorization profiles were constructed – one to represent the classes of Technology Solution Conference Sections, and one to represent the classes of Industry Solution Conference Sections. Rule-based categorization compares the knowledge base defined for a category with the knowledge base represented in the document to determine “goodness of fit” indicator. SAS technologies calculate goodness of fit metrics for all classes in a profile. A critical review of the metrics allows us to select one or more best fits. This presents a significant improvement over manual processing because it support cross-classification where appropriate, is objective and is verifiable.

The categorization results also surface a single document’s matching concepts, again across classes. This capability allows us to define a deep semantic representation of a document, and a representation of concepts across documents for a repository.

CONSTRUCTING THE SEMANTIC SOLUTIONS

For Method 2 the most important task was knowledge engineering of the Conference Section Model, and its high level categories. *CONFERENCE SECTIONS* is a hierarchical classification scheme with two persistent subclasses at the highest level - *TECHNOLOGY SOLUTIONS* and *INDUSTRY SOLUTIONS*. These two top level classes have significant differences in structure, different knowledge bases, and behavior. The two structures cannot effectively be modeled as one classification scheme, both using the same semantic approach and one consistent knowledge base. *TECHNOLOGY SOLUTIONS* (Figure 5) is a potentially deeper hierarchy. *INDUSTRY SOLUTIONS* (Figure 6), on the other hand, has only a single top level structure – resembling a flat classification scheme. For the purpose of this exploratory research, both models were constrained to a single level of categories.

CONFERENCE SECTION - TECHNOLOGY SOLUTIONS MODEL

Over the years, there has been some variation in the Technology Solutions class scheme. The variations derive from sections which have split, joined or been added as new trends. While it is true that each of these classes can be further broken down into subclasses, a consensus on what those classes might be does not yet exist. It was decided to limit our model to the ten classes defined in Figure 5.

- Advanced Programming
- Application Development
- Basic Programming
- Business Intelligence and Analytics
- Data Integration
- Data Mining
- Data Warehousing
- Reporting and Information Visualization
- Statistics and Data Analysis
- Systems Architecture and Administration

Figure 5: Sections – Technology Solutions Classification Scheme (*Top Level Classes Only*)

CONSTRUCTING THE CONFERENCE SECTION CATEGORIZATION RULES

The classification structure was constructed in the SAS® Categorization profiling client (Figure 6 – left screen panel). A deep knowledge base was then constructed to support rules for each of the classes. The Technology Sections are generally described with one or two paragraphs on the SUGI Conference websites. These descriptions belie the deep knowledge base behind each section and covered by the presentations. The challenge was to quickly and accurately build that knowledge base for each class. It was our assumption that that simply using the conference presentations would not be sufficient for representing the knowledge base for *TECHNOLOGY SOLUTIONS*. Much tacit knowledge is “hidden” in these Section descriptions – based on the SAS community’s deep knowledge base of the SAS® technologies.

The following resources were used to construct the knowledge base which fueled the rule-based categorization for *TECHNOLOGY SOLUTIONS*: SUGI Conference Proceedings Tables of Contents and Indexes, SUGI Master Index, derived list of Author supplied keywords, SAS® Technical Documentation Tables of Contents and Indexes, SASopedia classes and content on sasCommunity.org, and noun phrases extracted from the conference proceedings themselves. The knowledge bases for each of the Technical Sections were derived iteratively by:

1. Manually classifying author supplied keywords to Technical Sections;
2. Manually aligning Master Index entries to Technical Sections;
3. Reviewing and comparing Technical Section concepts to Technical Documentation Table of Contents' and Indexes;
4. Reviewing and comparing SASopedia classes on sasCommunity.org to Technical Section concepts;
5. Finally, augmenting the list of baseline concepts with extracted noun phrases.

The iterative approach provided a robust knowledge base for the rules-based categorizer. A sample extract of the rule base for Statistics and Data Analysis is presented in Figure 6. The core concepts defining this class numbered around 1,500. This class continues to be augmented. The method of deriving concepts and rules for the knowledge base is both stable and extensible. It is stable in its essential structure, and it is extensible in its concepts and rules. The structure can be updated annually as new concepts surface.

There were challenges involved in working with each of these sources.

- For example, the SUGI Master Index makes heavy use of extended phrases – often three to six words in length including both verb and noun phrases. Another drawback is that the entries are often artificially inverted and do not reflect how a person would refer to the concept – “applications, educational” rather than “educational applications”.
- The value of the Master Index is that it highlights significant coverage of concepts.
- Author supplied keywords on the other hand were often unique (generally have only one reference point to proceedings) and are of varied quality. Author supplied keywords are high risk – it is difficult for users to guess what keywords might have been used, without a controlled vocabulary. Author provided keywords were used to expand the concept level definition of the classes.
- Where Author supplied keywords are integrated into the *SECTION* knowledge base, though, they can add unique value.
- SAS® Technical reference materials provide a high quality source of concepts for the knowledge base – these are well written, well indexed, and have well-formed structures (Tables of Contents). These materials provide the highest quality input for creating the knowledge base.
- The challenge in using the technical reference materials is that there isn't a one-to-one correspondence between their focus and the Conference Sections.
- Extracted noun phrases are the best description of concepts treated in the conference presentations – however, they also must be aligned with the Conference Sections.
- Concepts derived from SASopedia or from the SAS community site are valuable but do not necessarily align with the conference presentations. They do contribute to and expand the coverage of the knowledge base, though.

The matching algorithm for this categorization profile configured for Frequency Based matching. This means that both the number of matches and the frequency of matches were considered in calculating the goodness of fit indicator for each class for each document.

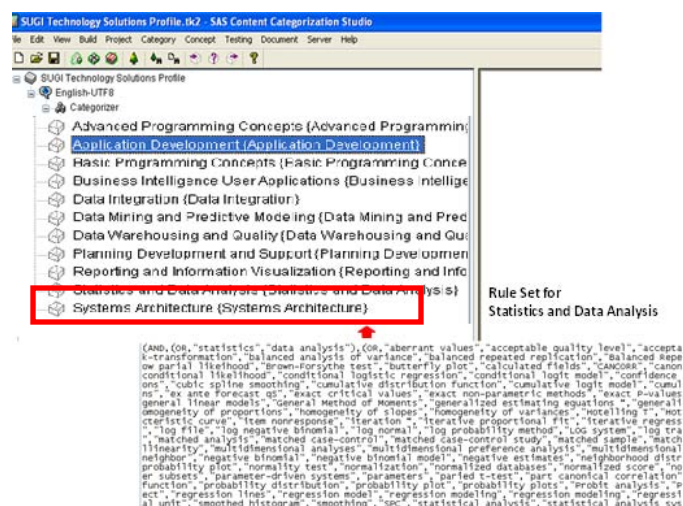


Figure 6: Categorization Profile for Sections – Technology Solutions

Each category was supported by a rich knowledge base of concepts (Table 7). As we would expect, the average rate of matching varied across categories for each document. Graphical representations of the clustered concepts for each category are presented below.

Table 7. Semantics Profiles for Industry Solutions

Conference Section	Total # Concepts in Category	Total Concepts Matched in Category	Average Match Rate Per Paper
Advanced Programming	845	217	34.88
Application Development	234	53	9.48
Basic programming	730	147	23.68
Business Intelligence and Analytics	114	13	1.08
Data Integration	602	115	19.72
Data Mining	415	25	2.04
Data Warehousing	521	43	4.04
Reporting and Information Visualization	589	90	12.84
Statistics and Data Analysis	910	79	10.56
<i>Total Concepts in Profile</i>	<i>4,960</i>		

CONSTRUCTING SECTIONS – INDUSTRY SOLUTIONS MODEL

The *INDUSTRY SOLUTIONS* is a simple, one-dimensional classification structure (Figure 7). While this structure has also varied over the years, the variations can be accommodated by including all of them.

- Banking and Financial Services
- Communications, Media, Entertainment and Travel
- Customer Intelligence
- Education Solutions
- Emerging Technologies
- Energy and Utilities Solutions
- Government Solutions
- Healthcare Providers and Insurers
- Insurance Solutions
- Internets Intranets and the Web Solutions
- Life Sciences Solutions
- Manufacturing Solutions
- Pharma Solutions
- Retail Solutions

Figure 7: Sections – Industry Solutions Classification Scheme

While the semantic solution is comparable for both *TECHNOLOGY SOLUTIONS* and *INDUSTRY SOLUTIONS*, the knowledge bases are very different. Whereas the knowledge base used for *SECTION - INDUSTRY SOLUTIONS* is a custom built from the rich SAS® content base, the knowledge base for *INDUSTRY SOLUTIONS* derived largely from the standard industry classification sources. The standard industry classification schemes are much more extensive than the coverage provided in the SAS® content. For example, there were far fewer *INDUSTRY SOLUTION* concepts in author supplied keywords than *TECHNOLOGY SOLUTION* concepts. Only 12.5% of the author supplied keywords related to *INDUSTRY SOLUTIONS* compared to 87.5% for Technology Solutions. To augment the knowledge base, we looked to some other industry standard profile descriptions of industries. An extract of the knowledge base for the Pharma Class of *INDUSTRY SOLUTIONS* is presented in Figure 8 below.

We deferred to standard definitions of Industry to construct the Industry Section profile. This worked well – the richness of concepts in each category (class) ranged between 500 and 3,000 concepts. These profiles should be made publicly available in the future for open and common use. To construct these profiles we also leveraged concept extraction technologies. This process took about two day's dedicated full-time effort because these class knowledge maps were readily available – we borrowed and adjusted.

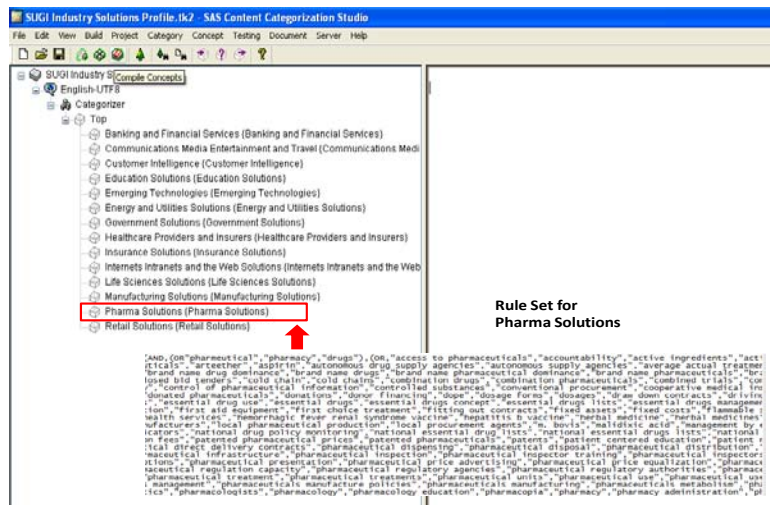


Figure 8: Categorization Profile for Sections – Industry Solutions

Each category was supported by a rich knowledge base of concepts (Table 8). As we would expect, the average rate of matching varied across categories for each document.

Table 8. Semantic Profiles Constructed for Industry Solutions Conference Sections

Conference Section	Total # Concepts in Category	Concepts Matched in the Category	Average Match Rate
Banking and Financial Services	3,015	53	6.36
Communication Media Entertainment and Travel	4,170	152	22.36
Education Solutions	13,666	83	14.54
Emerging Technologies	5,366	147	38.84
Energy and Utilities Solutions	4,353	18	1.8
Government Solutions	3,853	40	6.12
Healthcare Providers and Insurers	6,709	151	18.56
Insurance Solutions	2,616	26	3.92
Manufacturing Solutions	2,595	52	7.6
Pharma Solutions	992	30	3.44
<i>Total Concepts in Profile</i>	<i>47,335</i>		

The semantic profiles are applied to each paper individually to generate best fit categories (Conference Sections) and all of the matching concepts that could be used as keywords. Taken together these two sets of values can be used to construct new navigation and browse structures, and a richer search capability. We offer two views of the results of the content categorization work. The first view is that of the concept clusters that support the Conference Sections (Figures 9 through 17). The second view represents the categories and concepts aligned with individual documents (Figures 18 through 27).

CONCEPT CLUSTERS FOR TECHNOLOGY SOLUTION CONFERENCE SECTIONS

Figures 9 through 17 illustrate the concept clusters that were generated for each Technology Solutions Conference Section. The clusters include all of the concepts that were matched in the category. The size of the concept represents the intensity and frequency of the matching rates across documents. It also tells us something about the focus of the papers Vis a Vis sections over time. For example, four sections clearly are stronger than all others, specifically: Advanced Programming, Basic Programming, Reporting Design and Development, and Statistics and Data Analysis. The least populated Sections are Business Intelligence, Data Integration, Applications Development, Data Mining and Data Warehousing. This is not an intuitive result given the emphasis on these areas. There is a

[illegible]

15

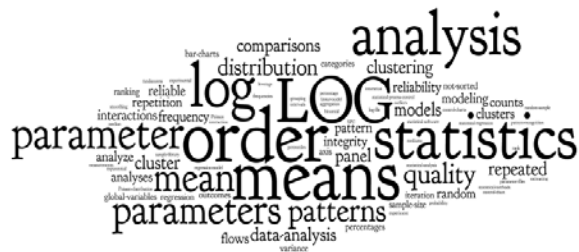


Figure 17: Statistics and Data Analysis Concept Cluster

CONCEPT CLUSTERS FOR INDUSTRY SOLUTION CONFERENCE SECTIONS

Figures 18 through 27 illustrate the concept clusters that were generated for each Industry Solutions Conference Section. The clusters include all of the concepts that were matched in each category. As with the Technology Solutions clusters, the size of the cluster represents the intensity and frequency of the matching rates across documents. It also tells us something about the focus of the papers vis a vis sections over time. What we see is a much sparser set of clusters for Industry Solutions. In fact, only three of the clusters represent a degree of density, including: Customer Management, Emerging Technology, and Health and Insurance Provider. In the case of both Customer Management and Emerging Technologies, there appears to be an affinity with general technology concepts. The results for these two clusters are not necessarily an indication of a strong Industry category. Health and Insurance Providers is a strong cluster, though. As we saw in the review of the Technology Solutions clusters, the challenge may be how the authors talk about or present the Industry Solutions. It may be that when the primary focus of the document is Insurance, the insurance concepts are much sparser than the use of technical terminology.



Figure 18: Banking and Finance Concepts



Figure 19: Customer Management Concept Cluster



Figure 20: Education Solutions Concept Cluster



Figure 21: Emerging Technology Concept Cluster





As illustrated in Table 1, the “Discovery” and the “Engineering “ methods have a lot in common with each other and are diverse enough to consider them different approaches to tackling the same problem. The results are very interesting in that the “Discovery” method’s top level categories match the level of keywords generated by the “Engineering” method.

Figures 30 and 31 illustrate the overall findings of the two methods of the content of the proceedings analyzed.



Figure 30: Full Concept Cluster for the “Discovery” method

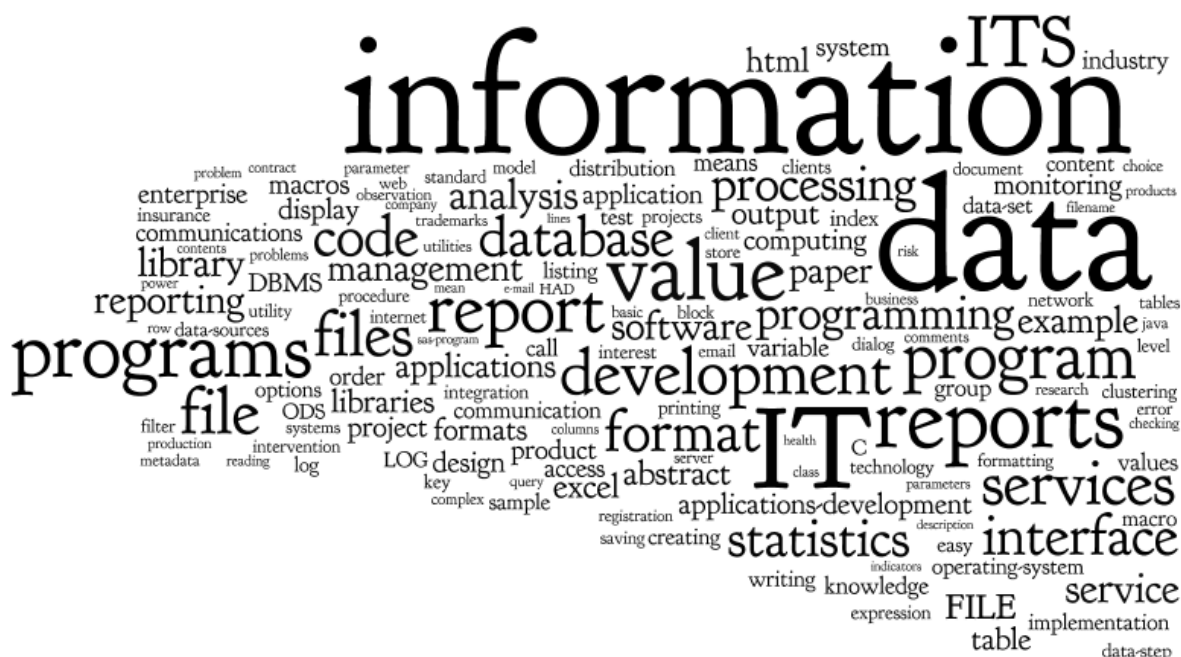


Figure 31: Full Concept Cluster for the “Engineering and Deep Semantic” method

19

It is clear that both approaches provide a great deal of synergy for working together. It may be possible to use the two to create a continuous feedback method between the two approaches which will allow the researcher to improve the automatic semantic understanding.

This effort represents the first stage of research of ongoing path of research into text analytics using SAS® technologies by the authors. The insight that the SAS® community will gain from this effort should improve the understanding of what content and topics are in the SUGI/SAS Global Forum proceedings. Unlike the shoemaker's children having no shoes, this effort is using the technology provided by SAS® to discover and improve the understanding of this proceedings content so that it can be used more effectively by the entire SAS® community.

REFERENCES

- Albright, Russell, Jake Bartlett, and David Bultman. 2007. "Making Web Results Relevant with SAS®" SAS Global Forum 2007. Cary, NC: SAS Institute Inc. Paper Available at <http://www2.sas.com/proceedings/forum2007/086-2007.pdf>
- Bedford, Denise A. D. (2012). Using Semantic Technologies to Analyze the Semantic Orientation of Religious Sermons – A Validation of the Early Work of McLaughlin, *Advances in the Study of Information and Religion*. October, 2012
- Bedford, Denise A. D. (2012). Visual Exploration of the Semantic Markers of Faith. *Advances in the Study of Information and Religion, ASIR 2012*.
- Bedford, Denise A. D. (2012) Enhancing the Precision of Geographical Tagging – Embedding Gazetteers in Semantic Analysis Technologies. Poster. *TKE 2012 Madrid Spain, June 19-22, 2012*
- Bedford, Denise A. D. and Karen F. Gracy (2012). Comparative Evaluation of Three Types of Semantic Distance Metrics – Implications for Use in Semantic Search, in Anthi Katsirikou and Christos Skiadas, eds. *Proceedings of the 4th International Conference on Qualitative and Quantitative Methods in Libraries, Limerick Ireland, May 23-25, 2012*
- Bedford, Denise A. D. and Karen F. Gracy (2011). Leveraging Semantic Analysis Technologies to Increase Effectiveness and Efficiency of Access to Information, *Qualitative and Quantitative Methods in Libraries October, 2012*
- Bedford, Denise A. D. (2011) Enabling Personal Knowledge Management with Collaborative and Semantic Technologies. *Bulletin of the American Society for Information Science* Volume 38, Number 2 December/January 2012, pp 32-39.
- Bedford, Denise A. D. and Karen F. Gracy (2011). Leveraging Semantic Analysis Technologies to Increase Effectiveness and Efficiency of Access to Information, in Anthi Katsirikou and Christos Skiadas, eds. *Proceedings of the International Conference on QQLM2011 Athens Greece – May 24 – 27, 2011*. (Accepted for publication).
- Park, Randi, Susan Waggoner and Denise Bedford (2010). Lessons learned in Content Architecture Harmonization and Metadata Models. *ASLIB Proceedings: New Information Perspectives, Vol. 62, No. 4/5, 387-405*.
- Bedford, Denise A. D. (2004). Using Concept Extraction, Categorization and Summarization Technologies to Fuel Semantic Search, in Michael E. D. Koenig and T. Kanti Srikantiah, eds. *Knowledge Management Lessons Learned: What Works and What Doesn't* Information Today, 2004.
- Bedford, Denise A. D. (2004). Designing Search Architectures to Support Knowledge Discovery and Management, in Michael E. D. Koenig and T. Kanti Srikantiah, eds. *Knowledge Management Lessons Learned: What Works and What Doesn't*. Information Today.
- Bedford, Denise A. D. and Jessica Gardner (2011). *Semantic Evaluation of the Performance Appraisal Feedback 2010 through Pilot Study*. Internal Report. World Bank Group, Washington DC, 2011.

Bedford, Denise A. and Richard W. La Valley 2011, "Improving Searchable Access to SAS® Global Forum Conference Papers Using Semantic Analysis Methods and Text Analytic Technologies," SAS Global Forum (SGF) 2011 Conference Proceedings, Kent State University and SAIC, Arlington, VA.

Cerrito, Patricia B. 2004. "Solutions to the Investigation of Healthcare Outcomes in Relationship to Healthcare Practice", Cary, NC: SAS Institute Inc. Paper available at <http://www2.sas.com/proceedings/sugi29/001-29.pdf>.

De Ville, Barry. 2006. "Text Mining with 'Holographic' Decision Tree Ensembles", Proceedings of the Thirty-first Annual SAS Users Group International Conference. Cary, NC: SAS Institute Inc. Paper available at www2.sas.com/proceedings/sugi31/072-31.pdf.

Jansen, Lex, SAS and JMP Paper Repository Website, <http://www.lexjansen.com/>.

La Valley, Rich and Nat Wooding, Nat (2011), "Using SAS® to Ease the Proofing of Messy Text", 2011 NorthEast SAS Users Group (NESUG) Conference and 2011 SouthEast SAS Users Group (SESUG) Conference, Strategic Technology Solutions, Herndon, VA, USA and Independent Consultant, Midlothian, VA.

La Valley, Rich, Lex Jansen and Kirk Lafler (2013) "Recovering SAS® User Group Proceedings for the SAS® Community", 2013 SAS Global Forum, Strategic Technology Solutions, Herndon, VA, [lexjansen.com](http://www.lexjansen.com), NJ, Software Intelligence Corporation, San Diego, CA.

LSA. Acronym for Latent Semantic Analysis. Both Latent Dirichlet Allocation (LDA) (<http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf>) and The SAS Text Miner Topic Node (<http://www.sas.com/resources/factsheet/text-miner-factsheet1.pdf>) are examples of LSA.

Sanders, A., & DeVault, C. 2004. "Using SAS® at SAS: The Mining of SAS Technical Support". Proceedings of the SUGI 29 Annual SAS Users Group International Conference. Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2004. Getting Started with SAS 9.1 Text Miner. Cary, NC SAS Institute Inc.

Zubair Shaik, Satish Garla, Goutam Chakraborty. 2012. SAS® Since 1976: An Application of Text Mining to Reveal Trends . Cary, NC: SAS Institute Inc. Paper 135-2012

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Barry de Ville
SAS Campus Drive
SAS Institute Inc.
Email: Barry.deVille@sas.com

Denise Bedford
Goodyear Professor of Knowledge Management
Kent State University, Ohio 44242-0001
Email: dbedfor3@kent.edu

Rich La Valley
Strategic Technology Solutions
Herndon, VA 20171
Email: rwlaivalley@gmail.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective 2-2013 companies.