Paper 162-2011

Improving Searchable Access to SAS® Global Forum Conference Papers Using Semantic Analysis Methods and Text Analytic Technologies

Denise Bedford, PhD, Goodyear Professor of Knowledge Management, Kent State University 44242-0001, USA

Richard W. La Valley, SAIC, Arlington, VA, USA

ABSTRACT

SAS® Global Forum has offered attendees the ability to search the contents of the papers being presented prior to the conference. This searchable access has parameters: title, author, abstract, keywords, and the conference section. Current semantic analysis technologies make it possible to leverage the natural language processing methods, organizational and domain knowledge, and the full text itself to achieve significant productivity gains–in extending access points for conference papers, generating more granular access to information, producing more objective descriptions and categorization of content. This paper presents experimental results on some of those gains.

KEY WORDS

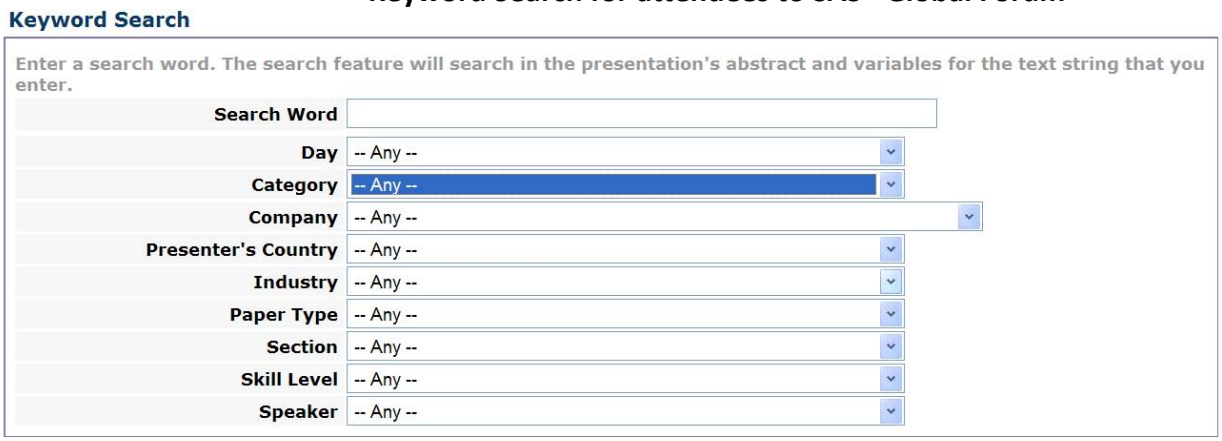
Topic Modeling, Automated Classification, Concept Extraction, Auto Summarization, Text Mining

INTRODUCTION

SAS® Global Forum conference has historically offered the attendees the ability to search the contents of the papers being presented prior to the conference. It has provided access to the abstract submitted, and has provided a keyword search function which allows the attendee to search by words in the abstracts submitted by the author and by various parameters such as day presented, category, company, country of presenter, industry, type of paper, section and the speaker. The interface for this tool can be found in Figure 1 below.

Figure 1

Keyword Search for attendees to SAS® Global Forum



Once the paper has been submitted and it is part of the conference proceedings, searchable access is supported by various parameters: title, author, abstract, key words provided by the

author, and the conference section to which the paper was assigned for presentation at the conference. This can be found when accessing the conference proceedings after the conference in web pages such as seen in Figure 2 and Figure 3.

Figure 2
Search for attendees to SAS® Global Forum 2010

This traditional approach to accessing conference information was efficient to produce and effective to search for many years. However, today users expect search to be at the concept level (i.e., more granular access), to be multi-faceted (many dimensions or parameters), and to be more precise (reduction of noise in result sets). Meeting these expectations with a manual indexing and classification approach is prohibitively expensive.

Current semantic analysis technologies make it possible to leverage the natural language processing methods, organizational and domain knowledge and the full text itself to achieve significant productivity gains – in extending access points for conference papers, generating more granular access to information, producing more objective descriptions and categorization of content. Productivity gains made possible by semantic analysis methods make it possible to now generate better and more precise access to content parts – a full paper, sections of the paper, paragraphs, or even the human written abstract.

TEXT MINING AND SEMANTIC ANALYSIS & SAS®

Text mining can be defined as the process of discovering information in large unstructured collections of documents and automatically identifying patterns and relationships in the text. Text mining builds off the insights from traditional data mining, natural language processing, machine learning, semantic web or Web 2.0, and information retrieval.[1] Text can be

Figure 3
Search pages from Lex Jansen's SUGI/SAS® Global Forum web access

represented in various levels such as lexical, syntactic and semantic. Lexical would include character, words, phrases, parts-of-speech, taxonomies and thesauri. Syntactic would include vector-space modeling, language modeling, full-parsing of the text, and cross modality. Semantic would include collaborative tagging such as in Web 2.0 technologies, templates and frames, and ontologies and first order theories.

SAS® provides this functionality by combining the structured data analysis and unstructured view in its SAS® Text Miner software and Teragram software capabilities. It combines natural language processing (NLP) and advances linguistic techniques to categorize the content of large volumes of multilingual content by parsing and analyzing the content for entities and events, creating metadata, assisting in the development of taxonomies and the generation of rules to define categories and concepts.

OVERVIEW OF THE PROBLEM AND RESEACH QUESTIONS

Problem Statement 1: Author generated titles are often key access points for conference presentations. In some cases, they may be the only access point. To what extent does the title of a conference presentation represent the content of the presentation or paper? It would be expected that the title of a conference presentation and the content itself would have very strong alignment of concepts.

***Research questions of interest**: Are titles reliable access points for conference proceedings? What level of variance is there between conceptual index terms for the title and for the content?*

Testing approach: An automated categorization profile for Conference Sections will be designed and implemented using the SAS® categorization suite. In the first test, conference paper titles and conference papers (full text) will be automatically categorized to Conference Sections. The results will be evaluated for consistency. In the second test, the conceptual keywords generated in the categorization process will be evaluated for consistency. We expect that the categorization and indexing that derives from the full conference papers will be better indicators of the content than the categorization and indexing results from titles.

Problem Statement 2: Manually generated abstracts will always be of higher quality than a programmatically extracted summary. However, programmatically generated abstracts may be more representative of the contents of the paper. Abstracts written by an author may be designed to attract attention to the presentation, or to gain acceptance in a submission process. The abstract may not always be a good representation of the conference presentation. Abstracts are the primary access point and are full-text indexed – for access by search engines that are used for SAS® Global Forum users.

***Research questions of interest**: Are abstracts in fact a good surrogate for the paper in search? When classified to the cross-conference classification scheme, do the abstracts and their papers classify consistently? When index terms (keywords) are generated for both the abstract and their papers, what percentage of the abstract terms are covered by the paper's terms?*

Testing approach: The same automated categorization profile for Conference Sections will be applied to abstracts. The results will be evaluated for consistency with those generated for the full text conference papers. Both the Conference Section values and the conceptual indexing terms generated will be evaluated. We expect that the categorization and indexing that derives form the full conference papers will be better indicators of the content than the categorization and indexing that results from the abstracts.

An additional test will be performed for abstracts. An automated extract will be generated for each conference paper, using the SAS® automated summarization capabilities. The automatically generated abstract will also be automatically categorized and compared first to the manually produced abstract, and then to the full conference paper.

Problem Statement 3: Additional access points beyond the five mentioned are important to researchers, particularly researchers who cannot attend the conference in person. Additional access points include: people referenced, institutions referenced, particular procedures used, economic sectors referenced, research type (experimental, applied, etc.), and specific technologies or applications mentioned.

***Research questions of interest**: Can we automatically generate these additional access points using semantic analysis technologies? Is the quality sufficiently reliable to use for search?*

Testing approach: Automated categorization and concept extraction profiles will be designed to address additional facets in the SAS® Global Forum Conference Proceedings search. The SAS® categorization suite technologies will be used to implement these profiles. Values for these additional access points will be developed leveraging the full-text conference papers. We expect that the new access points will be reliable and important sources of information for constructing the new search capabilities.

Problem Statement 4: Search results are most relevant when they take the searcher directly to the piece of information sought within the paper, not just to the whole paper.

***Research questions of interest**: Can the relevance of search results be improved by applying semantic analysis methods to parts or "chunks" of conference papers? What are the implications for deriving topic classes and indexing terms when components are automatically processed?*

Testing approach: In addition to applying the categorization methods to titles, abstracts and whole papers, the research considers how effectively the automated methods can be applied to "chunks" of conference papers. We expect that the automated approach will lead to improved access to content.

GENERAL METHODOLOGY

This methodology section provides an overview of the development and construction of the automated categorization and concept extraction profiles. The first step in designing any semantic solution is to analyze and model the challenge. The second step is to select a semantic solution which aligns with the challenge. Our challenge involves designing a semantic solution that generates values for the SAS® GLOBAL FORUM search facets – values that are as good as or better than those currently provided manually. For the automated process to perform as well as the manual process, the solution must include four components:

1. A deep and comprehensive knowledge base representation of the historical and current focus of each facet;
2. A semantic understanding of the conference presentation;
3. An appropriate semantic solution;
4. A method for determining either a goodness of fit or a confidence level.

The SAS® Global Forum Conference Papers search is a multifaceted search capability which covers: Category, Company, Country, Industry, Paper Type, Section, Skill Level, and Speaker. Each facet has a different structure, a different knowledge base, and a distinct behavior. One semantic strategy will not support all facets effectively. Success is entirely dependent upon understanding the behavior of the individual facets, knowing which knowledge sources to leverage, and selecting and configuring a suitable semantic solution. The solution set includes the semantic technologies supported by the SAS® categorization suite which includes five

semantic methods, including rule based concept extraction, grammar based concept extraction, rule based categorization, statistical categorization, and automated summarization.

SAS® SEMANTIC SOLUTIONS

The SAS® Categorization Suite offers five semantic solutions: (1) grammatical concept extraction; (2) rule-based concept extraction; (3) dynamic (statistical) categorization; (4) rule-based categorization; and (5) rule based sentence extraction. Several of these semantic solutions were used in this research: rule-based categorization, rule-based concept extraction and automated sentence extraction.

GRAMMAR BASED CONCEPT EXTRACTION

SAS® categorization suite allows the designer to define specific sets of grammatical units for extraction, using a standard tag sets.

RULE BASED CONCEPT EXTRACTION

Rule-based concept extraction allows the designer to define a set of rules or an explicit knowledge base for extraction. The designer can use regular expression (REGEX), define authority control lists, or use explicit values.

DYNAMIC OR STATISTICAL CLUSTERING

SAS® offers the standard statistical clustering capability based on Bayesian models. However, the categorization suite also supports the ability to "feed" a controlled set of values into the clustering engine, and to define boundaries and constraints for clusters.

RULE BASED CATEGORIZATION

Rule-based categorization involves defining the classification structure and constructing the knowledge base for each class. Rule-based categorization leverages either operator or frequency based matching algorithms to define a "goodness of fit" result for any document and for all classes in the structure. SAS®'s rule-based categorization technology produces a ranking of "goodness of fit" values to all classes in a scheme. This presents an improvement over manual processing which generates a single fit with no objective indication of "relative fit."

RULE BASED SUMMARIZATION

Most of the summarization tools available on the market today are either readers or extractors. Readers rely exclusively on frequency clustering and use the highest occurring keywords to automatically select sentences or fragments. Extractors use internal format representation, word and sentence weighting. Extractors produce higher quality extracts based on our anecdotal experience. SAS® 's rule-based summarization tools enable us to design rules into the extraction process.

SAS® 's automated summarization solution is a rule-driven pattern matching and sentence extraction technology. It extracts full sentences that are assembled to form a document surrogate or a "gist". The technology allows the designer to define rules and conditions for selecting sentences (specifically which key concepts should be used to select sentences, the position of these concepts in a sentence, the weighting of these concepts, which concepts to use to exclude sentences) and to specify how many sentences to select.

CHALLENGE OF THE FACETS

Our research results focus on a comparison of the manual and automated approaches to generating values for one of the facets – *CONFERENCE SECTIONS*. However, the research team began by modeling and testing all of the SAS® GLOBAL FORUM search facets. Figure 4 illustrates the semantic analysis strategy and the sources that would be used to automate each facet in the future.

It is clear from the initial tests that it is feasible to automate the capture of values for each of the facets using the SAS® Enterprise Content Categorization Suite. However, each facet requires a distinct profile and a distinct design effort. Analysis of the facets suggested that:

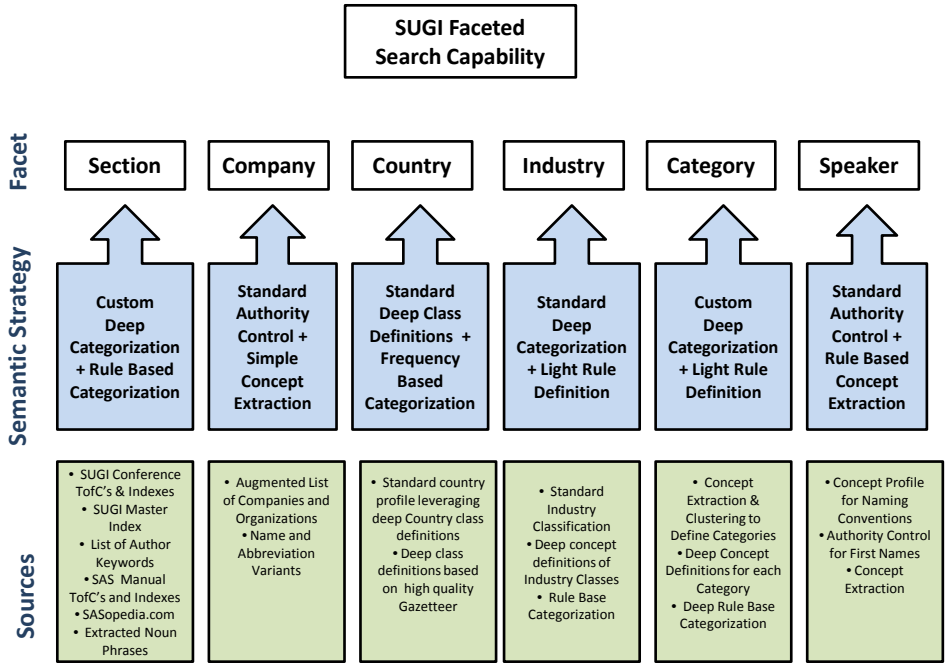
- Conference Section, having a strong predefined classification structure, could leverage deep rule-based categorization. However, a custom profile would need to be developed to align with the SAS® GLOBAL FORUM Conference Sections.
- *Category*, not having an agreed upon classification structure could first leverage a combination of concept extraction and statistical clustering to determine a good classification structure, and then leverage rule based categorization to automatically classify content;
- Country, having a universal classification structure, could leverage existing rule based categorization profiles to classify content;
- Industry, also having a good alignment with standard industry schemes, could leverage publicly available rule-based categorization profiles;
- Both Speaker and Company align most closely with concept extraction methods. Company leverages a standard authority controlled set of company and organization names, but in the form of an authority control file. Speaker leverages name patterns and contexts and increases precision by using an extensive authority file of international first names.

CONSTRUCTING THE SEMANTIC SOLUTIONS

A working semantic model of each facet was developed as a starting point. Each model took into consideration the human thought process used to generate the values, the type and number of values expected, the rules that a person uses to select values, and the knowledge sources a person would use. By way of example, a high level description of the modeling of four of the facets is presented below – *CONFERENCE SECTION*, *SPEAKER'S COUNTRY, SPEAKER, SPEAKER INSTITUTION*. As a starting point, though, we will focus on the facet which is most

relevant to the research results being reported to the SAS Global Forum 2011 Conference – *CONFERENCE SECTION*.

Figure 4
Semantic Strategy and Sources for SUGI/ SAS® GLOBAL FORUM Faceted Search



CONFERENCE SECTION MODEL

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.

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. It is also true that each of these classes can be further broken down into subclasses. However, because a consensus on what those classes might be does not yet exist, we decided to constrain our focus to those subclasses identified in Figure 5.

Rule-based categorization semantic technology was selected for this facet. 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.

Figure 5
Sections – Technology Solutions Classification Scheme
(Top-level Classes Only)

- Application Development
- Business Intelligence and Analytics
- Code Doctors and Coders' Corner
- Data Integration
- Data Mining and Text Analytics
- Hands-On Workshops
- Management
- Operations Research
- Planning and Support
- Programming
- Reporting and Information Visualization
- SAS Enterprise Guide – Implementation and Usage
- Social Media and Networking
- Statistics and Data Analysis
- Systems Architecture and Administration

CONSTRUCTING THE CONFERENCE SECTION CATEGORIZATION RULES

The Technology Sections are generally described with one or two paragraphs on the SAS® GLOBAL FORUM 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 to train the semantic technologies. It should be noted that simply using the conference presentations is not sufficient to represent 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:* SAS® GLOBAL FORUM conference proceedings tables of contents and indexes, SAS® GLOBAL FORUM 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:

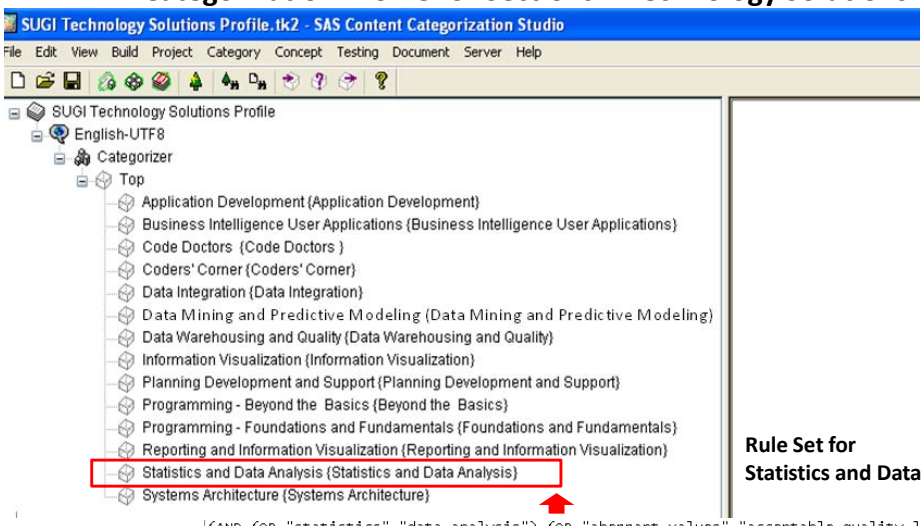
1. Manually classifying author supplied keywords to Technical Sections;
2. Manually aligning Master Index entries to Technical Sections;
3. Reviewing and comparing Technical Section concepts to SAS technical documentation table of contents' and indexes;
4. Reviewing and comparing SASopedia classes on 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.

The matching algorithm for this categorization profile was configured for Frequency-based matching.

Figure 6
Categorization Profile for Sections – Technology Solutions



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

- For example, the SAS® GLOBAL FORUM 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. However, they do contribute to and expand the coverage of the knowledge base.

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.

Figure 7
Sections – Industry Solutions Classification Scheme

- 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

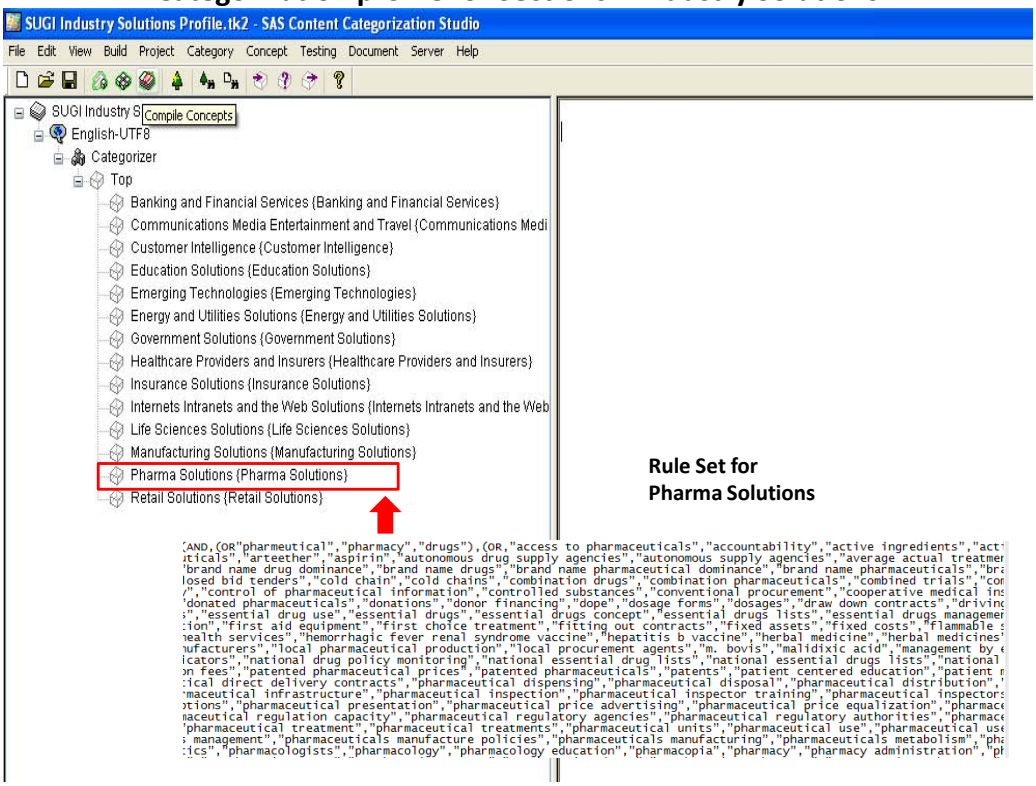
While the semantic solution is comparable for both *TECHNOLOGY SOLUTIONS* and *INDUSTRY SOLUTIONS*, the knowledge bases are very different. Whereas the knowledge base used for *SECTION - INDUSTRY SOLUTIONS* is custom built from the rich SAS® content base, the knowledge base for *INDUSTRY SOLUTIONS* derived 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 percent of the author supplied keywords related to *INDUSTRY SOLUTIONS* compared to 87.5 percent 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.

CONSTRUCTING A COUNTRY FACET MODEL

The manual process of defining the country of the presenter is also one of classification. The manual process involves identifying "country" cues and then assigning those cues to a value based on the person's deep knowledge of countries. The classification structure for *COUNTRY*

must be comprehensive rather than selective. As we can never predict with accuracy where speakers will come from, the knowledge base must cover all countries in the world. The classification process is essentially the same as it was for *SECTION* facets.

Figure 8.
Categorization profile for Sections – Industry Solutions



The semantic profile for Country is illustrated in Figure 9. The value of using an established knowledge source such as a public gazetteer is clear. An extract of the rule-base for the *COUNTRY* - China - is presented in Figure 9 below.

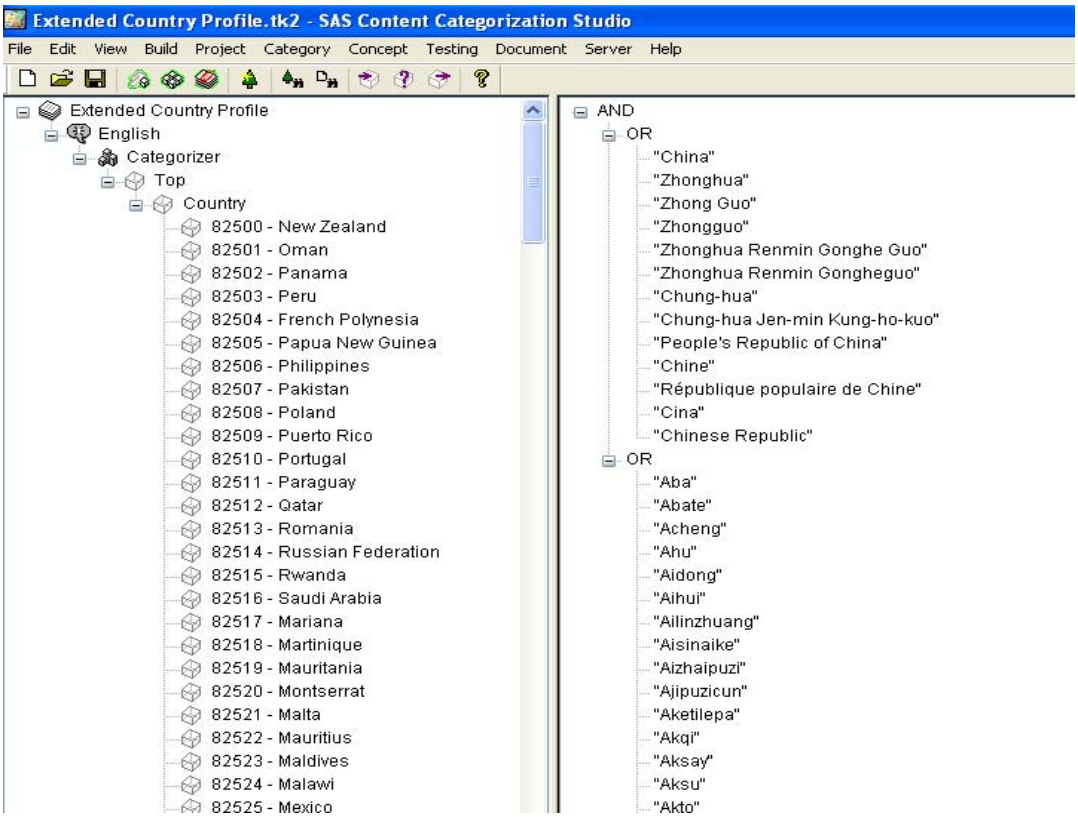
CONSTRUCTING THE SPEAKER FACET MODEL

The manual process of identifying speakers is based on knowledge of names, of name syntax and designations, and position in the paper. The knowledge base consists of a combination of an authority file of first names and rules representing contexts in which names may be found. The knowledge base of more than 32,000 first names was derived from web publicly-accessible sources of first names. Because names are universal, a standard profile – not customized for SAS® GLOBAL FORUM – could be used to automate this facet. An extract of the knowledge base for Speaker is presented in Figure 10 below.

Matching is binary – either a match is made or not made. Most attempts to identify names of people which are based solely on name syntax and position produce both false hits and false drops. The approach used in this research reduces both, but may tend to produce more false

hits depending on the nature of the first name. In order to use this approach for discovering authors, as opposed to any person referenced in the paper, pre-processing (i.e., chunking) of content is required – to ensure that it is applied only to the initial text of the document. Where a list of all people referenced in the profile is the goal, the profile is applied to the full document.

Figure 9
Categorization Profile for Country



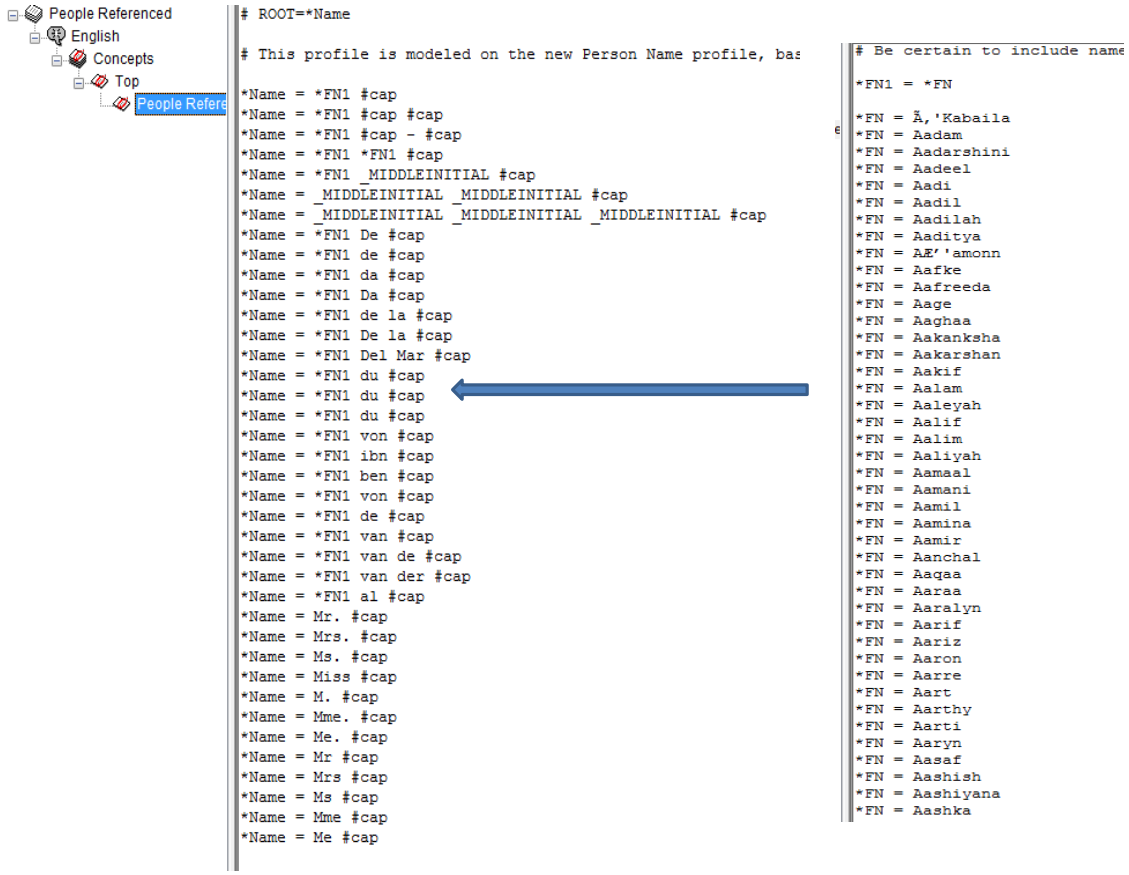
CONSTRUCTING THE COMPANY FACET MODEL

The manual process for identifying *COMPANY* is very similar to Speaker. The difference is the extensive authority control file that comprises the knowledge base. Because *COMPANY* names and abbreviations are universal, a standard profile could also be used for this facet. An extract of the knowledge base for *COMPANY* is presented in Figure 11 below.

SENTENCE EXTRACTION AND SUMMARIZATION

The summarization solution was used to automatically generate "gists" which were then (1) independently categorized to the two Conference Section profiles. Summarization rules were constructed for SAS® content. A sample rule set is presented in Figure 12.

Figure 10
Rule based Concept Extraction Profile with Authority Control File



CONCLUSION

The results of this research will be presented at SAS® Global Forum 2011. Initially, we started out to determine whether a better search capability could be developed using SAS® 's text mining capability. The research has provided a solid framework to structure the content in the SAS® Global Forum proceedings and indicates that there is a much richer search capability available to the SAS® community with the use of SAS®'s text mining tools. The knowledge organization approach used in this research for the SAS® content of the SAS® Global Forum proceedings is a solid foundation for the content and provides a new and improved way to look at the information contained in the papers submitted.

Figure 11
Rule Based Concept Extraction for Companies and Organization – With Authority Control

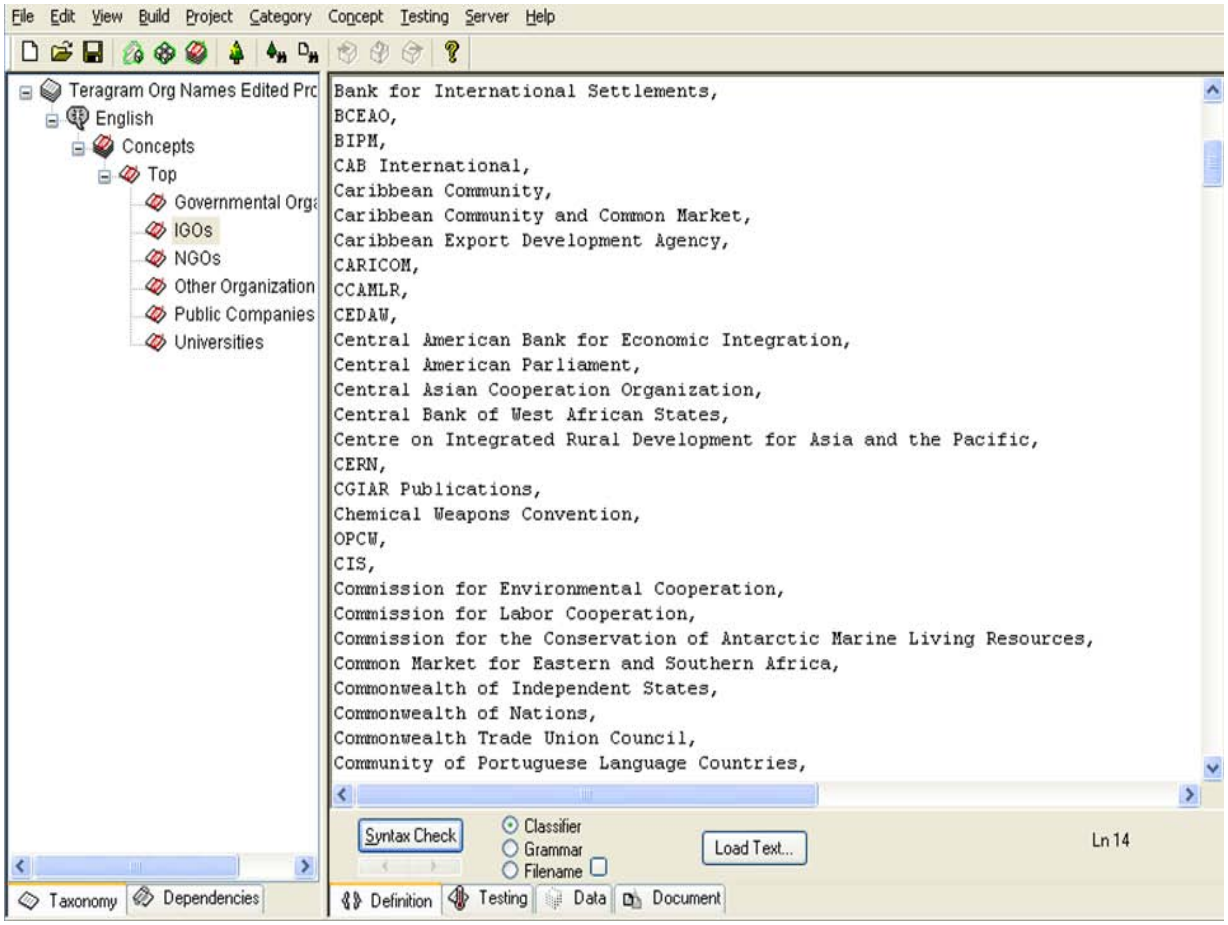


Figure 12
Sample Summarization Rules and Conditions

Code	Where would appear in the sentence	Inclusion Instructions	Syntax
5	anywhere in the sentence	It is likely not to be included	copyright/2004,5
9	anywhere in the sentence	Definitely not included	for/example,9
7	anywhere in the sentence	Definitely to be included	Was/the/solution,7
10	anywhere in the sentence	It is likely to be included	addressed/the/problem,10
2	anywhere in the sentence, followed by the second	It is likely to be included	evidence,2:suggested

1	beginning of the sentence	It is likely to be included	Result/was,1
6	beginning of the sentence	Definitely to be included	reporting/on,6
8	beginning of the sentence	Definitely not included	copyright/reserved,8
3	beginning of the sentence; only if the preceding sentence qualifies	It is likely to be included	however,3
4	beginning of the sentence; only if the preceding sentence qualifies	Definitely to be included	the/former,4

REFERENCES

Advanced Approaches in Analyzing Unstructured Data. New York, Cambridge University Press.

Denise A. D. Bedford, "Using Concept Extraction, Categorization and Summarization Technologies to Fuel Semantic Search," in Knowledge Management Lessons Learned eds. Michael E. D. Koenig and T. Kanti Srikantaiah, Information Today,2008

Denise A. D. Bedford, "Designing Search Architectures to Support Knowledge Discovery and Management," in Knowledge Management Lessons Learned, Michael E. D. Koenig and T. Kanti Srikantaiah eds., Information Today, 2008.

SAS® Content Categorizaiton Suite. Teragram TK240 User's Guide. Version: 5.1

Contact Information

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

Richard W. La Valley, SAIC
Director of Operations Analysis & Evaluation
4001 N. Fairfax Dr., Suite 700
Arlington, VA 22203
lavalleyr@saic.com