Paper 135-2012

SAS® Since 1976: An Application of Text Mining to Reveal Trends

Zubair Shaik, Satish Garla, Goutam Chakraborty,
Oklahoma State University, Stillwater, OK

ABSTRACT

Many organizations across the world have already realized the benefits of text mining to derive valuable insights from unstructured data. While text mining has been mainly used for information retrieval and text categorization, in recent years text mining is also being used for discovering trends in textual data. Given a set of documents with a time stamp, text mining can be used to identify trends of different topics that exist in the text and how they change over time. We apply Text Mining using SAS® Text Miner 4.3 to discover trends in the usage of SAS tools in various industries via analyzing all 8,429 abstracts published in SUGI/SAS Global Forum from 1976 to 2011. Results of our analysis clearly show a varying trend in the representation of various industries in the conference proceedings from decade to decade. We also observed a significant difference in the association of key concepts related to statistics or modeling during the four decades.

We show how %TMFILTER macro combined with PERL regular expressions can be used to extract required sections (such as abstract) of text from a large corpus of similar documents. A SAS macro developed for this research, %SAS1976, can be adopted to analyze papers published in any conference provided the conference papers are accessible in common formats such as .doc, .pdf, .txt, etc.

INTRODUCTION

Many organizations across the world have already realized the benefits of text mining by converting unstructured corpus of documents into structured data first and then applying data mining on the structured data to derive valuable insights. Firms with efficient algorithms for text mining have a competitive advantage over those who do not. Many researchers have published work related to applications of text mining in various domains. These applications mainly fall under the general categories of text categorization, information retrieval and measurement [1]. In recent years text mining is also being used for discovering trends in textual data. Given a set of documents with a time stamp, text mining can be used to identify trends of different topics that exist in the text. Natural Language Processing (NLP) is a widely used method for knowledge extraction. NLP is especially powerful in extracting predefined patterns (or existing knowledge) that can help in discovering trends from a research database [1].

Many professional conferences or forums are held every year across the globe. The total number of papers presented each year at all the conferences is likely in hundreds of thousands. While there are conferences which focus on only one specific field, there are many conferences which act as a platform for different fields. Unlike academic journals, which are heavily indexed and searchable, most conference papers are not indexed properly. In addition, there are many academic institutions that publish working papers and many consulting firms that publish white papers every year which are also not indexed. While it may be possible to use structured query to get a listing of papers published about a certain topic from indexed journals/conferences, it is virtually impossible to gain broad-based knowledge about the hundreds of topics that are presented or published during a time period and how such topics may have changed over time. In this paper we explain how text mining using SAS Text Miner® software can be used to identify trends in conference paper topics over time. As an example, we analyzed abstracts from all papers published each year at SUGI/SAS Global Forum from 1976 to 2011. Our approach can be followed to analyze papers published in any conference provided the conference papers are accessible in common formats such as .doc, .pdf, .txt, etc. We explore trending of topics in the published papers in at SUGI/SAS Global Forum across four decades as shown in Figure 1.

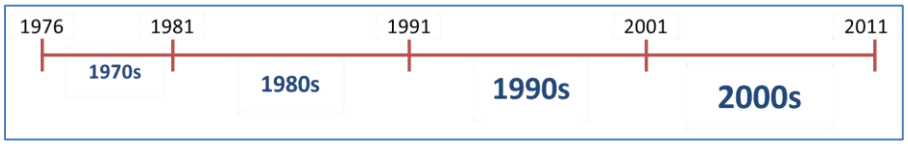

Figure 1 : Years divided into four decades in this analysis

Since the first SAS global conference started only in 1976, we have only five years of conference proceedings for the first decade. The number of papers per decade steadily increased from 1970s to 2000s as shown in Figure 2.

DATA

As mentioned earlier, the data we used in this study are abstracts of all SAS conference papers since 1976. Initially, we considered three types of textual data for text mining,

- ✓ Title of the paper
- ✓ Abstract of the paper
- ✓ Complete body of the paper

Title of the paper may not be a good input because it is often restricted in length and may not fully reflect the theme of the paper semantically. Considering the complete paper for analysis will likely add a lot of noise because a full paper can include tables, images, references, SAS programs, etc., which are problematic and may not add much value in text mining for topic extraction. We felt analyzing the abstract of a paper to be most appropriate since it captures detailed objective of a paper and does not contain extraneous items such as tables, images, etc.

We are thankful to SAS for making available SUGI/Global Forum proceedings for all the years for our research. We faced many challenges starting from downloading papers in PDF format from SAS website to preparing a final data set which contains only the abstract for each paper. We used %TMFILTER macro for preparing SAS data sets from a repository of SAS papers in .pdf and .txt format. We had to make some strategic choices to prepare the data sets. We explain in details our data preparation process in the appendix section of this paper. We created four data sets, one for each decade, with the number observations (i.e., paper abstracts) shown in figure 2.

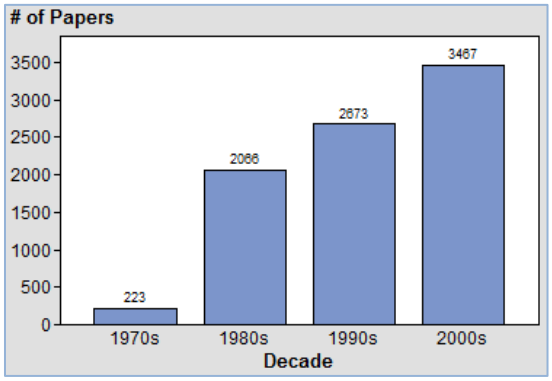

Figure 1 Number of paper abstracts in each Decade

TEXT MINING

We performed text mining on each of the four data sets separately. Text Mining starts with text parsing which identifies unique terms in the text variable and identifies parts of speech, entities, synonyms and punctuation [5]. The terms identified from text parsing are used to create a term-by-document matrix with terms as rows and documents as variables. A typical text mining problem has more terms than documents resulting in a sparse rectangular terms-by-document matrices [5]. Stop lists help in reducing the number of rows in the matrix by dropping some of the terms [1].

We used the same settings in the text miner node for analyzing each decade's data except the stop lists. Stop list is a dictionary of terms that are ignored in the analysis [5]. A standard stop list removes words such as "the, and, of, etc." However, a user can create custom stop lists for getting better text mining results [2]. For each decade, we created a different stop list that includes terms appearing in fewer than 10 documents as well as terms with highest frequencies. These terms are deemed as to not add any value to the analysis. Examples of such terms in the custom stop lists are SAS, SAS Institute, university, and so on. We also created a custom synonym data set using the terms extracted from all the four data sets. For example terms *Aviation* and *Airlines* were considered as synonyms for this research.

Even after using customized stop lists, in a corpus of several thousands of documents, the term-by-document matrix can contain hundreds and thousands of terms. It becomes computationally very difficult to analyze a matrix with high dimensional sparse data. Singular Value Decomposition (SVD) can be used to reduce the dimensionality by transforming the matrix into a lower dimensional and more compact form [5]. However, a careful decision needs to

be made on how many SVD high dimensions (k) to use. A high number for k can give better results, but high computing resources are required. It is customary to try different values for number of dimensions and compare the results [5]. As a general rule, smaller values of k (2 to 50) are useful for clustering, and larger values (30 to 200) are useful for prediction or classification [4]. In this study, by trial-and-errors, we settled on 25 and 50 as the maximum SVD dimensions for 1970s and all other decades respectively. Each term identified in text parsing is given a weight based on different criteria. The term weights help in identifying important terms [3]. The default setting for this property is Entropy. Using this setting, terms that appear more frequently will be weighted lower compared to terms that appear less frequently [2].

Clustering technique is used for text categorization. Using this technique documents are classified into groups such that documents within any one group are closely related and documents in different groups are not closely related [2]. The terms along with their weights are used for creating these groups. Each group or cluster is represented by a list of terms and those terms will appear in most of the documents within the group [3]. SAS Text Miner uses Expectation-Maximization algorithm for clustering. For all the four decades we first used 20 for the maximum number of clusters property and subsequently modified this property based on clarity of clustering results.

RESULTS AND DISCUSSION

SAS Text Miner's interactive window is a very useful tool for refining text mining results by changing various options. From this window we can see the clusters generated, term frequencies and weights and concept links. In the following section we first discuss the clusters and important concept links for each decade. In a subsequent section, we used the terms identified in each decade in discovering the trends.

CLUSTERS

Figures 3 -6 shows the clusters created by SAS Text Miner® .The theme of each cluster can be understood from the descriptive terms reported by SAS Text Miner. We can request text miner to report as many terms as needed in order to describe a cluster. In our analysis we used eight terms to describe a cluster. Our suggested cluster labels are based on the descriptive terms and corresponding documents.

Clusters

#	DESCRIPTIVE TERMS	FREQ
1	+ create, + tape, + store, + data set, + statement, + contain, + include, + clinical	17
2	+ linear model, + model, + general, + routine, + macro, + simple, + experiment, + statistical analysis	22
3	+ experiment, + measurement, + treatment, + factor, + table, + variance, + observation, + level	27
4	+ estimate, + square, + regression, + matrix, + variance, + model, + parameter, + compare	30
5	+ collection, + environmental, + research, + management, + requirement, + code, + function, + computer	13
6	+ package, + point, + measurement, + statistical analysis, + management, + general, + need, + system	23
7	+ plot, + produce, + option, + performance, + distribution, + univariate, + statement, + data set	15
8	+ system, + department, + report, + work, + clinical, + discussion, + management, information	51
9	+ record, + code, + produce, + contain, + data set, + develop, information, + problem	21

Clusters may be labeled as:

1. Storing clinical data
2. Statistical analysis using macro and linear models
3. Experimental methods of using SAS in measuring observations of different level
4. Regression & Model parameters estimate
5. Environmental research management
6. General need of statistical analysis package
7. Options in plots
8. Clinical Information management
9. Develop code to solve problems and produce information

Figure 2: Clusters created by SAS Text Miner® and their labels for the decade of 1970s

Clusters

# ▲	DESCRIPTIVE TERMS	FREQ
1	+ regression, + experimental, + measurement, + matrix, + estimate, + variance, + model, + square	259
2	+ update, + retrieve, + interface, + database, + datum management, + clinical study, + requirement, + develop	151
3	+ macro, + macro statement, + ad-hoc, + program, + valuable, + generate, + tool, + function	94
4	+ screen, + software, + command, + color graphics, + environment, + workstation, + discussion, + application	377
5	+ dataset, + update, + screen, + generate, + program, + database, + macro, + produce	79
6	+ chart, + statistical analysis, + clinical study, + study, + color graphics, + valuable, + numerical, + produce	549
7	+ consultant, + technique, + solution, + workstation, + application, + develop, + efficiently, + requirement	298
8	+ forecast, + energy, + model, + market, + company, + estimate, + develop, + method	63
9	+ program, + language, + ad-hoc, + efficiently, + generate, + load, + function, + valuable	196

Clusters may be labeled as:

1. Experimental measurement models
2. Developments required in data management
3. Macro and ad-hoc programming
4. Graphics and workstation environments
5. Macro Applications
6. Statistical analysis methods in clinical study
7. Consultant Techniques and application development
8. Applications market forecast
9. Effective usage of different programming

Figure 3: Clusters created by SAS Text Miner® and their labels for the decade of

Clusters

# ▲	DESCRIPTIVE TERMS	FREQ
1	+ program, + macro, + statement, + programmer, + parameter, + automate, + store, + display	166
2	+ method, + analysis, statistical, + plot, + parameter, + calculate, + statistic, + regression	400
3	+ train, + budget, + medium, + teach, + student, + path, + specific, + identify	48
4	+ graphics capability, + presentation, + good, + produce, + programmer, + program, + problem, + display	289
5	+ query, + performance, + access, + store, + database, + system, + database market, + environment	332
6	+ operate, + access, + application, + system, + platform, + server, + web, + operate system	296
7	+ survey, + read, + record, + merge, + statement, + process, + program, + technique	284
8	+ control, + development, + interactive mode, + develop, + graphics capability, + screen, + application, + object	375
9	+ treatment, + clinical, + drug application, + patient, + analysis, + research, statistical, + method	214
10	+ service, + industry, + forecast, + business, + model, + customer, + database market, + improvement	267

Clusters may be labeled as:

1. Automation using programming
2. Methods of statistical analysis
3. Training and education
4. Graphical capabilities and problem identification
5. Performance in database market
6. Web server and operating systems
7. Data preparation and cleaning methods
8. Applications of interactive mode and graphical capabilities
9. Patent analysis and research application in clinical study
10. Business model and database market

Figure 4: Clusters created by SAS Text Miner® and their labels for the decade of 1190s

Clusters

#	DESCRIPTIVE TERMS	FREQ
1	+ industrial, + measurement, + research, + clinical study, + visualization, + analyze, + method, statistical	499
2	+ performance, + interface, + web application, + server, + administrator, + applications, + environment, + client	574
3	+ market, + product, + registration, + business, + risk management, + retail, + service names, + register	408
4	+ regression, + selection, + logistic regression, + logistic, + estimation, + model, statistical, + method	308
5	+ government, + quality, + business, management, + organization, + resource, + system, + technology	418
6	+ function, + regular, + regular expression, + manipulation, + programmer, + feature, current, + technique	209
7	+ programmer, + merge, + technique, + feature, + method, + interface, + structural equation, + application	635
8	+ macro, + programmer, + automatic, + reference, + customize, + function, + problem, + purpose	290
9	+ custom, + appearance, + plot, + statistical graphics, + symbol, + customize, + scatter plot, + modify	133

Clusters may be labeled as:

1. Applications in different Industries
2. Web applications and client server administration
3. Risk Management and retail business
4. Statistical models and regression
5. Business applications and government
6. Programming features and technique
7. Applications in clinical research
8. Customized functioning and problem solving with macro
9. Statistical applications in economic research

Figure 5 : Clusters created by SAS Text Miner® and their labels for the decade of 2000s

While the clusters seem to change from decade to decade, the cluster definitions are very sensitive to the terms that are considered for clustering. Dropping or adding terms often drastically change the cluster patterns. Identifying the best text clusters is time consuming and need trial-and-errors.

CONCEPT LINKS

Concept links help in understanding the relationship between words identified in the documents [2]. Many people find concept links intuitive and easy to understand since they are presented as graphs with connecting terms. The width of the line between the centered term and a concept link represents how closely the terms are associated. A thicker line indicates a closer association. Not all terms have concept links [5]. Concept links can also be used to identify trends across four decades. As an example, we show below the concept links for the term ‘forecast’ in all four decades.

In 1970s the concept ‘forecast’ is mainly associated with terms such as relationship, time series, estimate, etc. which shows that papers in these decades are perhaps about what forecasting is and how SAS can be used for solving a forecasting problems. In 1980’s we begin see terms such as inventory, business , sale etc. showing up with forecast suggesting that the papers are focusing on application areas related to those terms. Whereas in 1990s we see new terms such as railroad, economic research suggesting ne application areas of papers published. Finally in 2000s we see new terms such as merchandise, retail and financial showing up in concept links suggesting even more specialized application areas investigated in those papers.

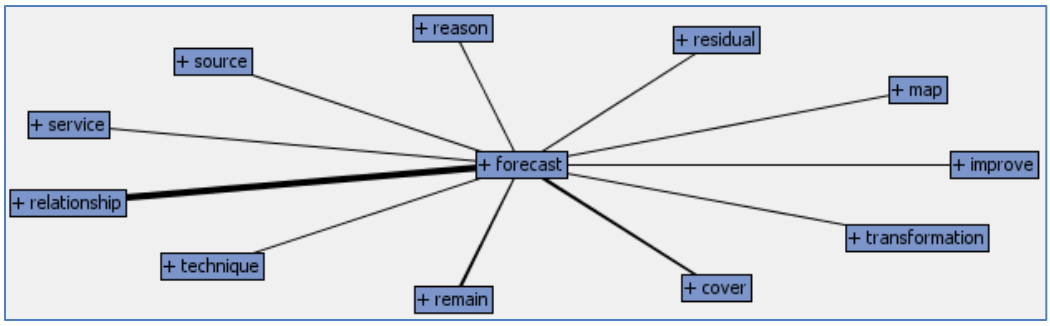

Figure 6: Concept links for the term 'Forecast' in 1970's

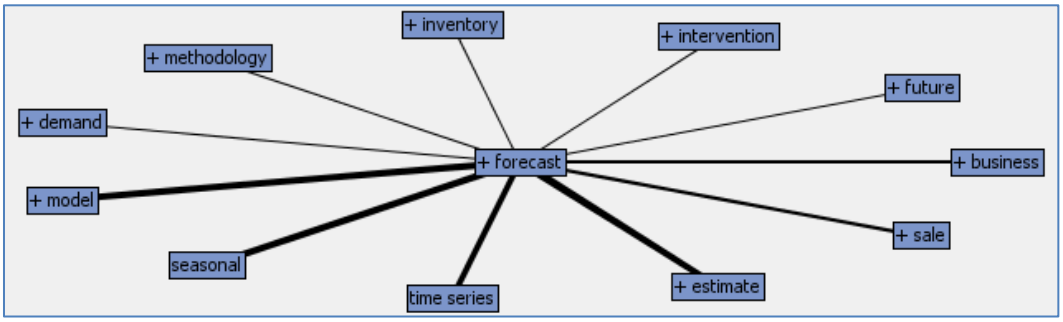

Figure 7: Concept links for the term 'Forecast' in 1980's

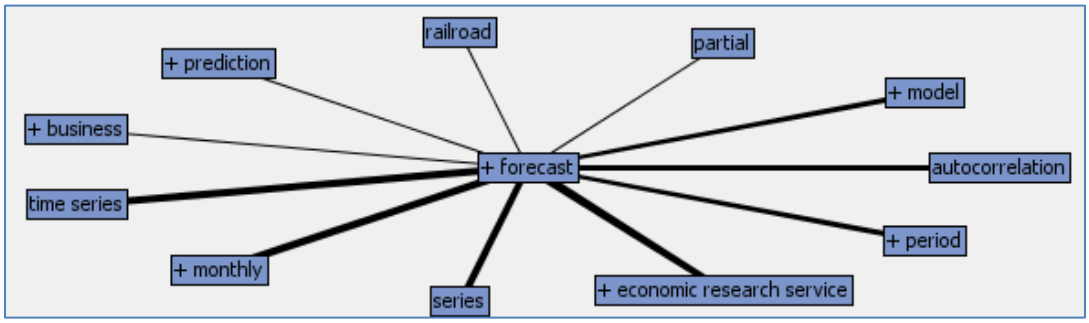

Figure 8: Concept links for the term 'Forecast' in 1990's

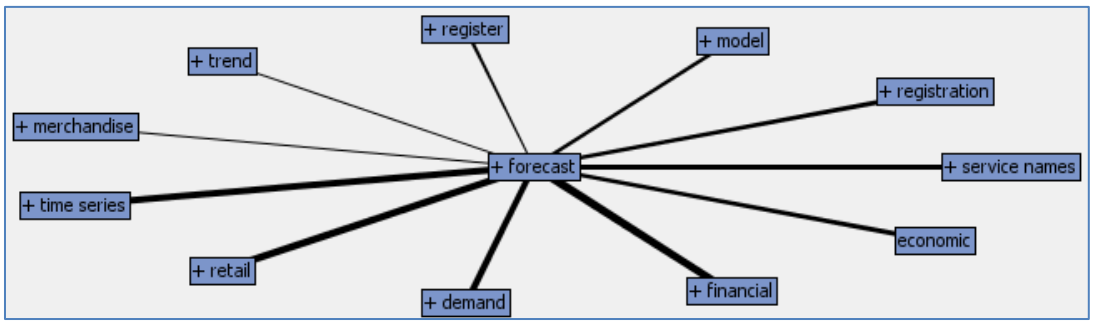

Figure 9: Concept links for the term 'Forecast' in 2000's

TRENDS

We used terms as topics to identify trends across the four decades. Terms related to a specific industry are treated as synonyms. The appearances of industry specific terms across the four decades are used to understand the trend for that industry. A frequency value of 'n' for a term means that particular term was mentioned in the abstracts of 'n'

distinct conference papers. Figure 10 shows the percentage of papers contributed for six of the top industries across all the four decades.

Most of the papers in 1970s (24.42%) were presented on Clinical/Healthcare/ Pharmaceutical industries followed by Financial (5%), Agriculture (4%) and Biotechnology (4%) domains. The same trend was not observed in the following years. Percentage of papers published on Manufacturing and Government gradually increased from 1970s through 1990s. However, in 2000s there was relatively fewer numbers of papers published in those industries.

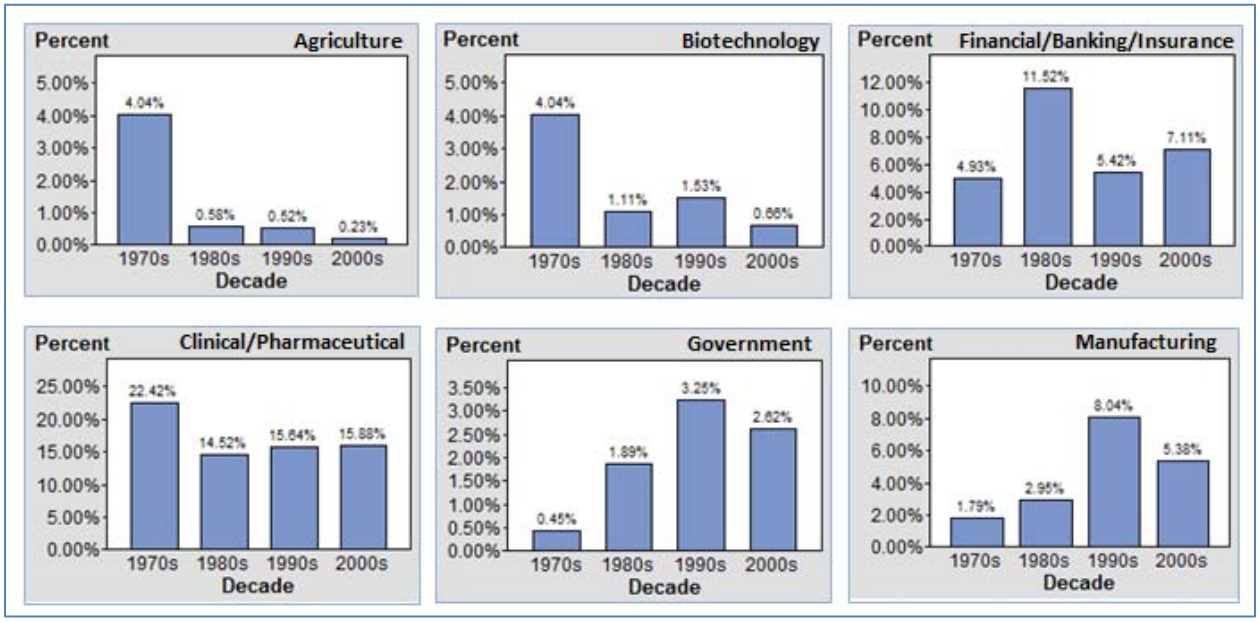

Figure 10: Percentage of papers contributed for top six industries across the four decades

In the same way, we can analyze the trend for other specific industries. Figure 12 shows the trend plot of few other industries with much less representation in the papers compared to the large scale representation of the industries as shown in Figure 11.

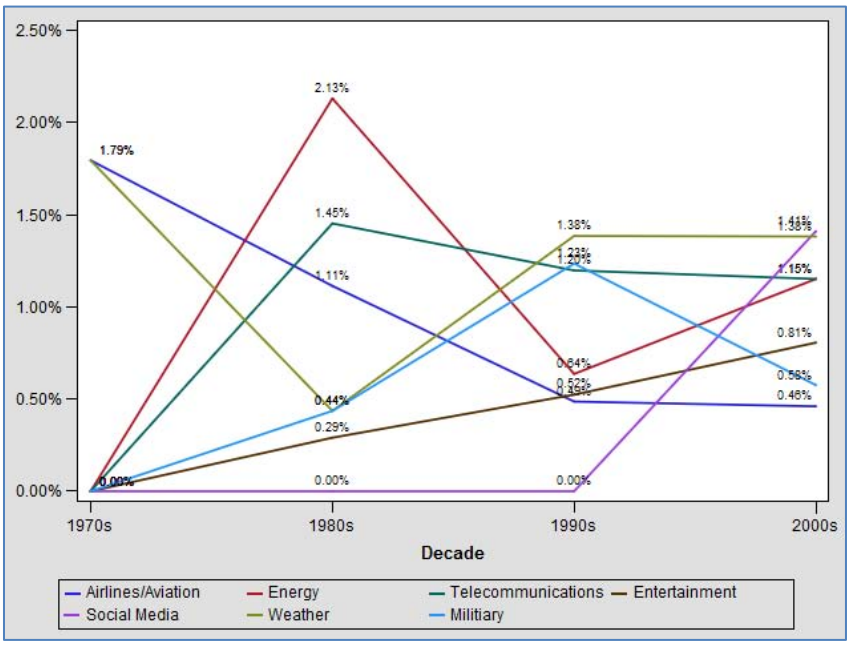

Figure 11: Trend plot of other important industries across the four decades

It is widely known that the growth of social media happened in 2000s and hence we can expect papers published on this industry during the decade 2000s. It is gratifying to observe that trend in the plot (Figure 11). The plot also shows the trend for, telecommunications, aviation, entertainment, weather, military and energy industry.

CONCLUSION

We applied text mining to figure out trends in research topics related to various industries in SAS/SUGI conference papers. Our analysis can be easily extended to find trends in research topics related to different methods or technology or procedure. A similar approach can also be used to analyze the many conference proceedings corpus that is available with various organizations across the globe. Text Mining has tremendous potential in identifying trending topics during a period of time. A user can create programs using %TMFILTER macro and PERL regular expressions that give unlimited capabilities in processing textual data. PERL regular expressions in SAS can be used for cleaning textual data. A SAS macro that we developed for the purpose of this research can be easily adopted by other researchers to investigate similar terms/topics trend in their corpus of documents.

APPENDIX

DATA PREPARATION

We are thankful to SAS and lexjansen.com for making all SUGI/SAS Global Forum proceedings available online. We downloaded all the papers from www.sascommunity.org and www.lexjasen.com. In the website sascommunity.org, the proceedings for the years 1976 till 1996 and 2004 till 2011 are available as direct downloadable links (PDF Files) and for the years 1997 till 2003 they are available in a protected PDF file format which makes it difficult to download all the papers in a single click. However, lexjansen.com website can be used for these years to download such files directly.

All the proceedings are available in portable document format (PDF). Downloading each proceeding year by year starting 1976 till 2011 is a tedious and time consuming process. We overcame this challenge using HTTrack, free software available from www.httack.com, which can download files from any website to your machine in one click. The data is saved on your machine in the same way it is stored on the website server.

Once we have all the papers in PDF format, we used %TMFILTER macro to create data sets from PDF Files. Given a folder with a set of PDF files, %TMFILTER macro can create a SAS data set with each PDF file as a row in the data set. However in our study we do not need the entire PDF contents for text mining. We only need the abstract for a paper. We used the NUMBYTES= option in the macro to extract only the first 2,000 bytes of text from each PDF file. The number 2,000 is an approximation since in most cases an abstract doesn't account for more than 1,500 bytes of text data. After extracting the first 2,000 bytes of text the main challenge is to identify the exact abstract. We used the key words ABSTRACT and INTRODUCTION to identify the exact text corresponding to an abstract. This was done using PERL regular expressions.

However we met with challenges in applying this abstract extraction technique for all the decades. We found that,

- The conference proceedings in 1970s, 1980s did not follow a particular format, i.e., all the proceedings did not contain the key words 'ABSTRACT' and 'INTRODUCTION'.
- In 1990s, there were documents with only the key word ABSTRACT without the keyword INTROCUTION and vice versa.

We solved the above challenges using a SAS Macro program (%SAS1976) that we developed for this research. The macro can extract text based on above mentioned conditions. The macro contains a loop mechanism in which the %TMFILTER macro is invoked on an iterative basis. This macro should be used separately for each decade. In every iteration, %TMFILTER macro first creates a data set of papers (first 2,000 bytes) for a year and then uses PERL regular expression to extract the required text based on four different conditions as described below.

Any of the below four situations are possible:

- The text contains both key words 'Abstract' and 'Introduction'
- The text contains only key word 'Abstract'
- The text contains only key word 'Introduction'
- The text do not contain both key words

The macro follows the below mentioned strategy in each of the above mentioned possible cases.

		INTRODUCTION	
		NO	YES
ABSTRACT	NO	All text between 150 and 1500 bytes is extracted	2000 bytes of text after the key word INRODUCTION is extracted
	YES	2000 bytes of text after the key word ABSTRACT is extracted	Text between the key words ABSTRACT and INTRODUCTION is extracted

There are exceptions where both the key words ABSTRACT and INTRODUCTION exist in the text but the keyword 'Introduction' appears before the key word 'Abstract' or the keyword 'Introduction' is the first few words of the Abstract. In such cases we made sure the length of the text extracted is at least 500 bytes. Otherwise the key word 'Introduction' is ignored and the remaining text is extracted. This technique is applied for each year's data set and in the end all the individual data sets for a decade are merged.

```
/*************************************************************************
Macro Name:        sas1976
Purpose:           Macro to create SAS dataset from large text corpus

How it Works:      This macro uses %TMFILTER macro to convert the .txt/.pdf
                   files present in a single folder (in this case all SUGI
                   papers for one year) to SAS dataset. PERL regular
                   expressions are used to fetch the text between the keywords
                   ABSTRACT and INTRODUCTION. And the datasets are
                   concatenated to form a bigger data set representing a
                   decade. We executed this macro four types with year values
                   changed to create four different datasets for four decades

Requirements:      Create a library and enter the path for the dir= option in
                   %TMFILTER that points to your data source. The text files
                   should be arranges in folders. One folder for each year
                   with folder name as the year number

*************************************************************************/

libname SAS1976 'D:\Zubair\EDU\Spring 2011\SAS1976';

%macro sas1976;
%let year=1976;

%do year=1976 %to 1980;
%tmfilter(dataset=sas1976.SAS&year.,
                    dir=D:\Zubair\EDU\Spring 2011\SAS1976\&year.,
              numbytes=2000);
data SAS_&year.;
length text $ 2000;
length name $ 100;
      set sas1976.SAS&year.;
if _n_=1 then do;
            retain pattern1 pattern2;
            pattern1 = PRXPARSE ("/ (Abstract) | (bstract) /i");
            pattern2 = PRXPARSE ("/Introduction/i");
      end;
position1=prxmatch(pattern1,text);
position2=prxmatch(pattern2,text);
```

```
/*If intro comes before abstract;*/
If position1 gt position2
then position2=0;
if position2-position1 lt 500
then position2=0;
if position1 gt 0 and position2 gt 0
then
      text= substr(text,position1+9,position2-position1-9);
else if position1 eq 0 and position2 eq 0
then
      text=text;
else
      text= substr(text,150,1500);
keep name text;
run;
proc append base=SAS1976.sas1976_1980 data=SAS_&year. force;
run;
%end;
%mend;
%sas1976;
```

REFERENCES

[1] Miller, W.T., (2005). Data and Text Mining-A Business Applications Approach. Pearson Pentice Hall.
[2] Cerrito, B.P., (2006). Introduction to Data Mining using SAS® Enterprise Miner. SAS Publishing.
[3] Battioui, C. (2008). A Text Miner analysis to compare internet and medline information about allergy medications. SAS Regional Conference.
[4] Sanders, A., & DeVault, C. (2004). Using SAS® at SAS: The Mining of SAS Technical Support. SUGI 29.
[5] "Introduction to Text Miner." In "SAS Enterprise Miner Help." SAS Enterprise Miner 6.2 . SAS Institute Inc., Cary, NC.
[6] SAS Institute Inc. 2010. Text Analytics with SAS ® Text Miner Course Notes. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

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

Zubair Shaik, Oklahoma State University, Stillwater OK, Email: zubairs@okstate.edu

Zubair Shaik is a Master's student in Management Information Systems at Oklahoma State University. He has two years of professional experience as Software Engineer. He is SAS Certified Base Programmer for SAS® 9 and Certified Predictive Modeler Using SAS® Enterprise Miner 6.1®

Satish Garla, Oklahoma State University, Stillwater OK, Email: satish.garla@okstate.edu

Satish Garla is a Master's student in Management Information Systems at Oklahoma State University. He has three years of professional experience as Oracle CRM Consultant. He is SAS Certified Advanced Programmer for SAS® 9 and Certified Predictive Modeler Using SAS® Enterprise Miner 6.1

Dr. Goutam Chakraborty, Oklahoma State University, Stillwater OK, Email: goutam.chakraborty@ okstate.edu

Goutam Chakraborty is a professor of marketing and founder of SAS and OSU data mining certificate program at Oklahoma State University. He has published in many journals such as *Journal of Interactive Marketing*, *Journal of Advertising Research*, *Journal of Advertising*, *Journal of Business Research*, etc. He chaired the national conference for direct marketing educators in 2004 and 2005 and co-chaired the M2007 data mining conference. He is also a Business Knowledge Series instructor for SAS.