Paper 1900-2018

From unstructured text to the data warehouse: customer support at the University of North Texas

Nick Evangelopoulos, University of North Texas

ABSTRACT

Traditional business intelligence uses a data warehouse to generate reports that increase organizational knowledge. In the current big data environment, organizational data include large collections of unstructured documents, especially in the domain of customer support. However, a formal process of expanding the traditional dimensional model to include elements that are derived from such collections is often missing. In this presentation, we provide a case study from IT Shared Services (ITSS) division at the University of North Texas (UNT). As part of the UNT System's ServiceNow initiative, we examine a collection of ITSS' service work-order data to take unstructured text to the data warehouse. Going beyond traditional reporting elements, such as service requests by time, or request category by department, we show how text analytics help uncover the hidden dimension of service topic. Based on this dimension, derived facts, such as certain service tickets addressing certain topics, are added to the data warehouse. These uncovered elements represent a part of organizational knowledge that would otherwise remain undetected, and can be used by decision makers to improve customer support and address service issues.

INTRODUCTION

Organizational knowledge exists in two main forms: tacit, the form that traditionally has resided in people's minds, and explicit, the form that is typically stored and managed with the help of information systems. Faced with the challenge of retaining tacit knowledge even after its human holders have departed, organizations engage in a process of converting tacit knowledge to explicit and explicit to tacit following the so-called spiral model (Nonaka and Takeuchi, 1995).

As individuals live, work, interact, and learn together within communities of practice, a collective tacit knowledge emerges (Kabir and Carayiannis, 2013). This is knowledge that cannot be acquired from individuals: its conversion to explicit knowledge requires recording and analysis of community interactions. The representation of this domain takes two forms: network relationships that reflect the topology of interactions among the community members, and collections of documents that reflect the information content they exchange as they accumulate their knowledge.

This paper is organized as follows. We start with a dimensional model that represents tacit organizational knowledge in the context of a service provider. We continue with a case study that illustrates the process of expanding the data warehouse design by adding new facts and dimensions, derived from the analysis of unstructured text. We conclude with a discussion of challenges encountered and lessons learned.

A DIMENSIONAL MODEL FOR SERVICE REQUESTS

The most common IT architecture that supports efficient generation of business intelligence reports in the knowledge organization is the *dimensional model*, or *star schema* (e.g., Adamson and Venerable, 1998, pp. 9–17; Kimball and Ross, 2002, p. 16). Figure 1 shows an example dimensional model for service requests. The fact table includes a description text, as the main content-related attribute, and a number of foreign key attributes that reference the dimension tables that surround the fact table. These dimension tables include a department, a location, and a time dimension.

Occasionally, the information stored in the data warehouse does not stop at information that is obvious at the point of original data capture, but extends to include attributes and dimensions that are produced later, as a result of some analytic process. Custom fact tables and custom dimensions are then included in the dimensional model (Adamson and Venerable, 1998, p. 20). Adopting database management terminology, we refer to these new facts as *derived facts* and their corresponding dimensions as *derived dimensions*. Figure 2 shows the topic dimension as an addition to the dimensional model that was obtained through the analytic process of topic extraction. Since topic extraction is a process that requires the full set of documents in order to be performed, the topic dimension represents tacit knowledge: Indeed, a certain document's topical coverage is not explicitly given to begin with, but involved employees are expected to share an understanding about it. However, with the help of text analytics, topics can be extracted, and documents can be associated with them. Therefore, text analytics contributes to a conversion of elements of tacit knowledge to explicit. For a more detailed discussion, see Evangelopoulos, Shakeri and Bennett (2018).

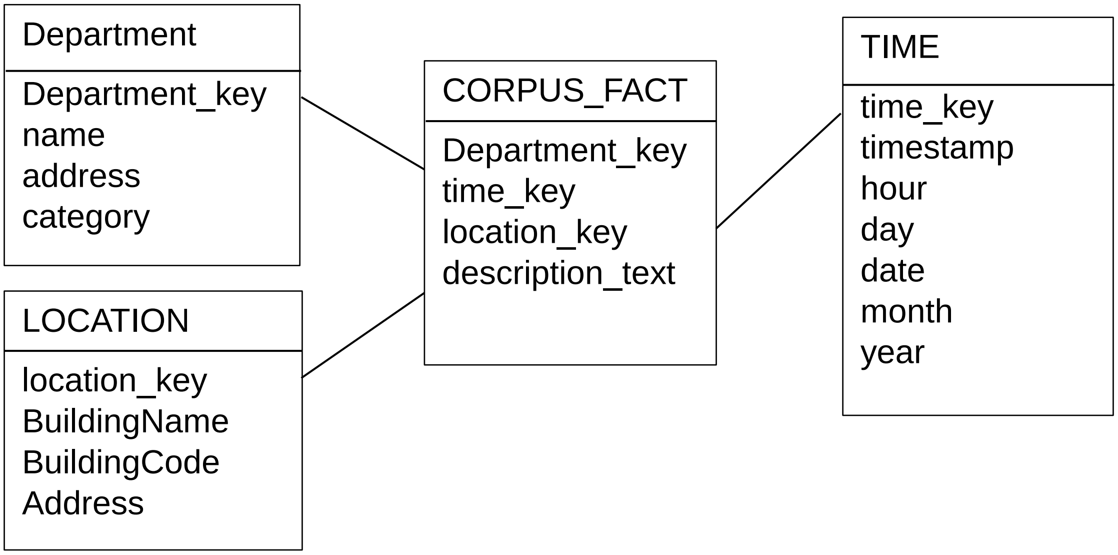


Figure 1. Dimensional model for a corpus of service requests

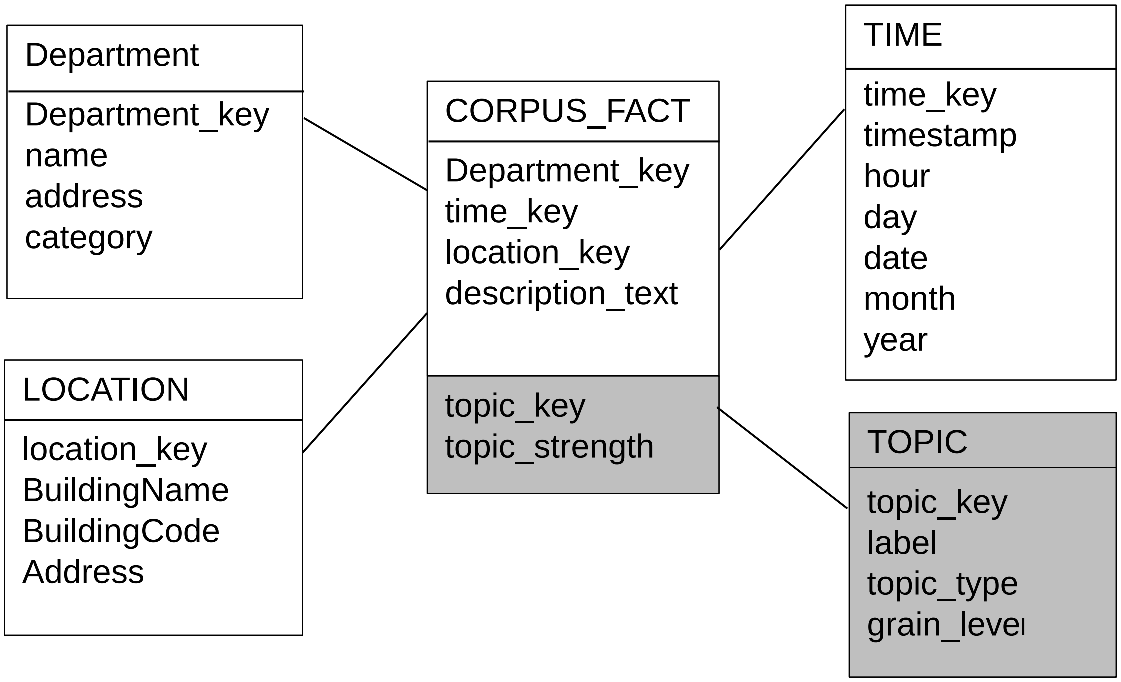


Figure 2. Expanded dimensional model with topic as a derived dimension

In the next section we introduce a case study that provides a context where the ideas presented here can be illustrated in an applied form.

CASE STUDY: IT SUPPORT AT THE UNIVERSITY OF NORTH TEXAS

BACKGROUND AND DATA COLLECTION

The Information Technology Shared Services division (https://itss.untsystem.edu/services) serves the UNT System's community of 5,000 employees and 40,000 students. ITSS supports the entire student life cycle – from recruiting, admissions and financial aid to advising and degree audits – helping to ensure their success. ITSS provides more than 60 enterprise IT shared services to supported its institutions. On average, 150 incidents are created during each business day, covering various IT-related issues. However, currently, incidents are not classified properly. Even though the current service work order management system has an incident category provision, most incidents are not categorized, and most of the rest are associated with a generic "inquiry" category. ServiceNow Kingston, the division's work order management system, allows for easy classification of incidents. However, the categories need to be customized to match the actual service requests. The questions addressed in this case study are:

- How can ITSS provide more efficient service ticket routing?
- How can ITSS identify appropriate incident types and categories?
- How can ITSS design an improved knowledge base to better support insightful and actionable business intelligence?

For the purposes of this case study, 9691 ITSS incidents that occurred from May 2016 to September 2017 were exported from the ServiceNow platform. Each incident record included information on the date and time of the incident's creation, resolution, and closing, the requesting department within the organization, the location, and a description of the incident in unstructured text form. Some incident descriptions were typed by ITSS associates because the requesting customer walked in or made a phone call, some others were submitted using a web interface, and others were submitted via e-mail.

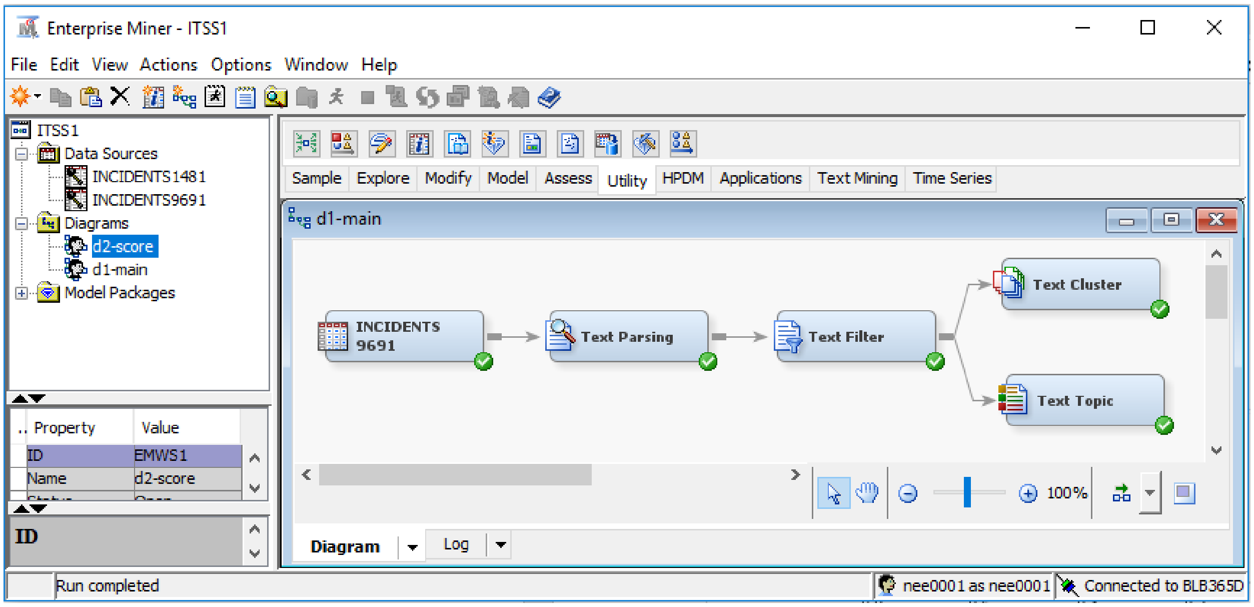
TEXT CLEANING OPERATIONS

In addition to information that was relevant to the nature of the incident included in the service ticket, Incident descriptions contained a large number of e-mail addresses, web site URLs, names, and phone numbers. The e-mailed incident descriptions were particularly problematic because they included a large e-mail trail with as many as ten or twenty rounds of back-and-forth responses, follow ups, and forwarded messages. Most of the cleaning operations were performed using regular expressions in the statistical programming environment R.

TOPIC EXTRACTION IN SAS® TEXT MINER

In order to understand the semantic structure of the body of the incident descriptions, we performed topic extraction in SAS® Text Miner. We start by creating a new analysis project in SAS® Enterprise Miner™. We created a new SAS Library, imported the incident description data file as a Data Source, opened a new analysis diagram canvas, selected the Text Mining tab on the toolset space and built the diagram shown on Display 1, using Enterprise Miner/Text Miner version 14.2. The analysis diagram consists of a Data Source node (INCIDENTS9691), a Text Parsing node, a Text Filter node, a Text Cluster node, and a Text Topic node. For details on the use of the Text Mining nodes, see Chakraborty et al. (2013, pp. 122–130). In the Text Parsing node settings, a list of synonyms that can be customized, is included. The text Parsing node includes a list of terms to be excluded from further analysis, known as a stop list. This can also be customized. For example, the stop list can include placeholder phrases that replaced original e-mail addresses or URLs, so that topic extraction results disregard language use patterns that revolve around the mentioning of e-mail addresses or web site URLs, and focus on overarching high-level concepts.

In the Weightings section of the Text Filter node settings, we selected Frequency Weighting = None, and Term Weight = Inverse Document Frequency. In the Term Filters section of the Text Filter node settings, we set Minimum Number of Documents = 4, so that terms that appear in less than four documents in the collection are filtered out.



Property	Value
ID	EMWS1
Name	d2-score

Display 1. Analysis Diagram for Topic Extraction in SAS Text Miner

Our modeling intention was to extract topics that describe the entire document collection, not to produce clusters that segment the documents. The purpose of the Text Cluster node in the analysis diagram was simply to produce some statistics that can help us determine an appropriate number of topics to extract. Since topic extraction is based on the dimensionality reduction operation of singular value decomposition, an examination of the eigenvalues can help determine an appropriate number of topics. Eigenvalues can be produced by squaring the singular values, which are one of the products of singular value decomposition. The Text Topic node does not provide access to the singular values, but the Text Cluster node does. Therefore, we performed text clustering and we browsed through the variables generated by the system in the background, to find the singular values. The scree plot of eigenvalues (see Figure 3) suggests 3 and 9 as candidate high-level dimensionalities. In order to explore the semantic content of the incident descriptions at a reasonable level of detail, we selected *k* = 9. Therefore, in the Learned Topics section of the Text Topic node settings, set Number of Multi-term Topics = 9.

Next, we ran the analysis diagram from the Text Topic and viewed the Results. Text Miner produces a summary of the extracted topics that lists, for each topic, a few characteristic terms that help identify the topic, and the number of related documents. With the help of the characteristic terms and some browsing through associated documents, the 9 topics were labeled as shown in Table 1. The incidents address phishing attacks, issues with UNT e-mail accounts, issues with classroom equipment, requests for software, budget process errors, compromised accounts, hardware requests, password resets, and office licenses. The main data warehouse design was adjusted to accommodate the topics as incident document facts.

SCORING NEW DOCUMENTS

At the conclusion of the analysis described in the previous sections, ITSS management decided to move forward with the implementation of a redesigned data entry form for service tickets, which incorporated a pull-down menu that made available the 9 topics as pre-defined incident categories. The idea was to have the users select one of the 9 topics (or an “other” category), and hopefully improve the service ticket routing. However, the final production system included some modification in the incident categories. One was eliminated because it was considered controversial, and two were split into sub-categories. In the period December 2017 to February 2018, 1,481 incidents were entered into the system. These incidents were scored by SAS EM into one or more of the original 9 topics, using the analysis diagram shown in Display 2.

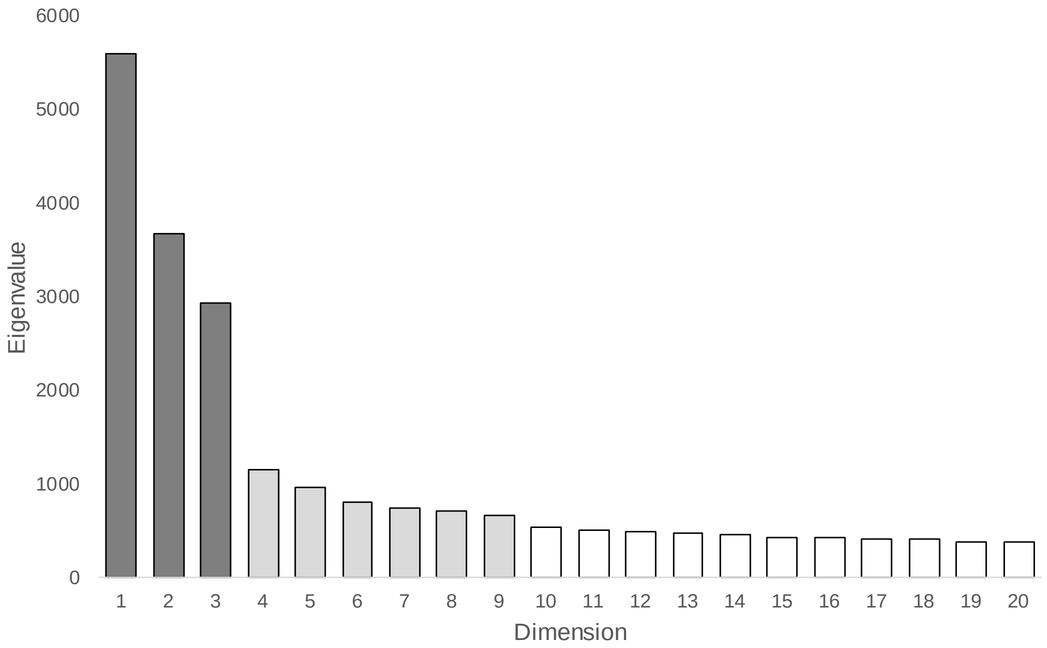
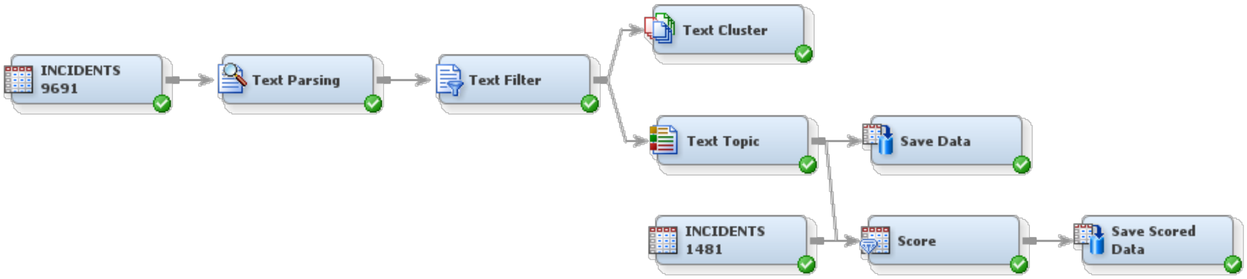


Figure 3. Scree Plot of Eigenvalues Produced by Singular Value Decomposition

Topic ID	Topic label	Related terms
T1	Phishing attack	+attack,+phish,+email account,+account,+block
T2	UNT email account	+address,email,+student,+email address,unt
T3	Classroom equipment	projector,readiness,equipment,+room,retractable
T4	Software provision	adobe,java,firefox,routine,firefox java
T5	Budget process errors	+process,+budget,+id,ps,+journal
T6	Compromised account	+password,+account,+update,+user,suspicious
T7	Computing device request	+install,+printer,pc,+laptop,+computer
T8	Password resets & Golden	+password,+access,golden,+customer,+reset
T9	Office license	+license,+error,+office,+mailbox,+log

Table 1. Extracted Topics and Corresponding Topic Labels



Display 3. Analysis Diagram for Scoring New Documents Against the Extracted Topics

CORRESPONDENCE ANALYSIS IN SAS/STAT®

After extracting topics from the incident descriptions and adding that information to the data warehouse, the next step was to produce some more complex queries that would offer insights on the interaction between topics and other dimensions. Information related to such interaction produces cross-tabulations, which can be visualized using corresponding analysis. Correspondence analysis emerged in the 1930s, as a method related to Singular Value Decomposition (SVD) and principal component analysis (Hirschfeld, 1935). The method projects a set of contingency table categories into a smaller number of orthogonal categories, the principal components. In preparation for correspondence analysis, the data set saved by the Save Data node needs to be restructured. Table 2 shows the original structure of the file, where an indicator variable is created for each extracted topic. Only the first three rows of data are shown. The original format needs to be converted into a long categorical data format, where a categorical variable *Topic* holds the topic value of each observation. This is shown in Table 2. Again, only the first three rows of data are shown. Documents with incident numbers 36750 and 43450 are associated with topics T4 and T7, and with location L42. The structured query language (SQL) code that was executed at the data warehouse to produce Table 3 is as follows:

```
Select INCIDENT_FACT.IncidentNumber, LOCATION.Location, INCIDENT_FACT.Topic
from LOCATION, INCIDENT_FACT
Where LOCATION.LocKey = INCIDENT_FACT.LocKey
Order By INDENT_FACT.IncidentNumber;
```

T1	T2	T3	T4	T5	T6	T7	T8	T9	_DOC_	Location
1	0	0	0	0	0	0	1	0	1	L42
0	0	0	0	0	1	0	0	0	2	L42
0	0	0	0	0	0	1	0	0	3	L42

Table 4. Format of the Data Set Produced by the Save Data Node

IncidentNumber	Location	Topic
INC0036750	L42	T4
INC0036750	L42	T7
INC0043450	L42	T7

Table 5. Long Categorical Data Format

The data set in long categorical data format is now ready for correspondence analysis. The Base SAS code shown below assumes the presence of a data file called ConsumerTopics.csv in location d:\temp, and uses PROC CORRESP to perform correspondence analysis:

```
proc import out= work.topics
    datafile= "d:\temp\IncidentTopics.csv"
    dbms=csv replace;
    getnames=Yes;
proc corresp all data=topics outc=Coord;
    tables Topic, Location;
run;
```

Execution of the above code produces a contingency table, chi-square information, profiles, and all results of the correspondence analysis. The OUTC= option creates an output coordinate data set. The TABLES statement specifies *Topic* and *Location* as the row and column categorical variables, respectively. Figure 4 presents the correspondence analysis map, which is a Topic-Attribute Map. The 9-by-54 contingency table is projected onto a two-dimensional space. Since the contingency table has nine rows, the complete information does not get projected on two dimensions. The two principal components account for 71.77% of the variability in the contingency table. The correspondence map provides an interesting insight: locations L6, L13, L15, L18, L25, L48, and L53, are all highly associated with topic T5 (budget errors). Therefore, these are seven locations where the frequency of budget errors is high. This association is shown on Figure 5 as a dashed oval shape.

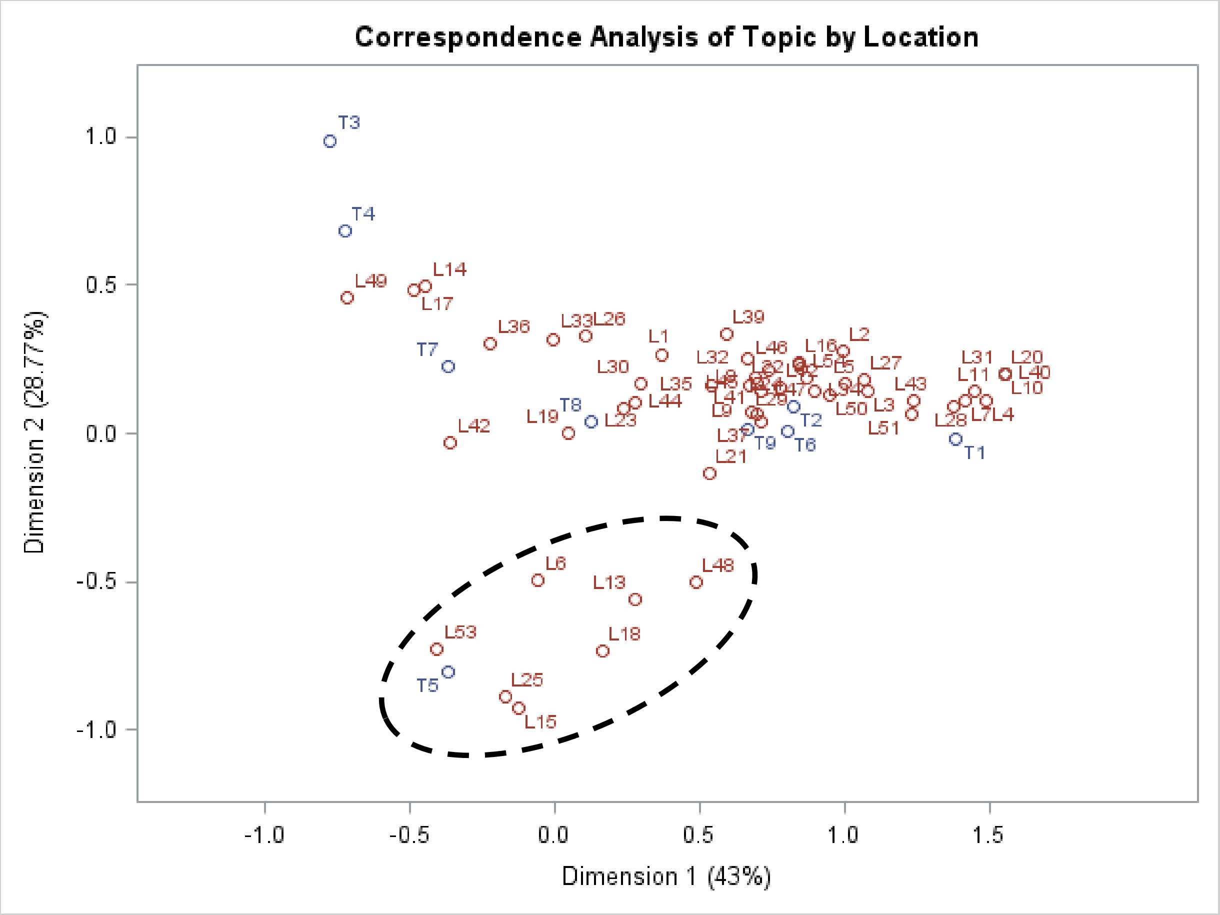


Figure 6. Correspondence Map Showing Topics and Locations

NEW DOCUMENT SCORING VERSUS USER CLASSIFICATION

Since we have the 1,481 new incident documents scored by SAS EM and also categorized by the users at the point of service ticket initiation, we can compare the two classifications. Table 6 shows the two distributions of document counts and Figure 7 visualizes them in the form of a bar chart. We observe that, while topics T1, T2, T3, T4, and T6, have similar distributions, topics T7, T8, and T9 exhibit dramatic differences.

Topic	Scored by SASEM	Selected by user
T1	18	11
T2	87	90
T3	0	1
T4	4	6
T5	115	N/A
T6	59	87
T7	98	3
T8	87	11
T9	14	1

Table 7. Classification of new documents by SAS EM scoring and user selection

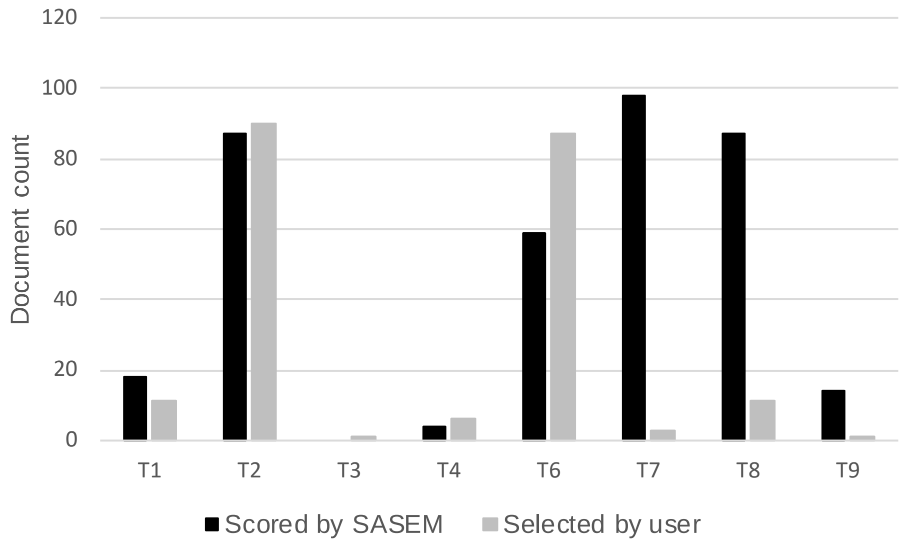


Figure 8. Comparison of the two classification distributions across the topics

DISCUSSION

Before the extraction of topics from the set of the original 9,691 incident descriptions, the data warehouse design focused on overall trends in service request volume. The data warehouse was used to produce reports that showed service request volumes by the day of the week, by the month of the year, by department, by location, etc. After the extraction of topics, the data warehouse model was adjusted to accommodate these topics as service request attributes. At that point, the reports were able to explore the interaction between topics and the rest of the star schema dimensions. The case study described one of them, addressing the interaction of the topic and location dimensions. The topic-attribute map (correspondence map) helped us make associations between certain locations and certain topics. Topic-Attribute Maps can be useful tools in helping the analyst gain actionable insights. They can also help the analyst tell a nice story, and, at the same time, visualize it. In this paper, the documents described customer service requests, and their attributes included the location where they originated within the organization. The analytic and visualization approach that was illustrated here can be extended to work with associations between topics and a variety of attributes that declare the origin or the purpose of the documents in the data set. For example, if the documents are news stories, the attribute of interest could be the news source, the locality mentioned in the news, or the time the events took place. If the documents describe leadership initiatives, the attributes can be organizational leaders or organizational units. If the documents describe customer reaction/feedback, the attributes can be products, services, or brands. For some additional discussion on the use of correspondence analysis maps as a visualization tool for concept spaces, see Evangelopoulos (2016a, 2016b).

Once we moved from insight to action, organizational elements were introduced to the analytic process. The results of topic extraction suggested a set of 9 topics as IT service request categories, but those were only partially implemented. One topic (budget errors) had a nature that reflected a retrospective assessment. Making the determination that budget errors have occurred was considered inappropriate for the stage of initiation of a service ticket. Certain topics were deemed too broad and were broken down into more detailed categories (e.g., computing devices were broken down to printer, PC, laptop, and network), and others proved too hard for the users to understand (e.g. password resets).

CONCLUSION

The case study presented in this paper extracts high-level concepts from collections of documents and visualizes them using Topic-Attribute Maps. Overall, the case study provides support for the following conclusions:

- SAS Enterprise Miner is an effective tool in extracting high-level concepts from collections of unstructured text documents.
- Correspondence analysis provides a way to produce informative output that facilitates comparison among document attributes that can be linked to brands, providers, leaders, products, services, topics, sentiment levels, locations, or time periods
- In a real production environment, when one moves from insight to action, the analyst is not always at liberty to implement the interventions suggested by the analysis. Other factors must be taken into account, including the history of the organizational unit, the culture, the leaders' initiatives, and the users' sense-making process, just to name a few.

REFERENCES

Adamson, C., and Venerable, M. 1998. *Data warehouse design solutions*. New York: John Wiley & Sons.

Chakraborty, G., M. Pagolu and S. Garla. 2013. *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS®*. Cary, NC: SAS Institute.

Evangelopoulos, N. 2016a. "Thematic orientation of the ISJ within a semantic space of IS research." *Information Systems Journal*, 26:39–46.

Evangelopoulos, N. 2016b. "Text Analytics and Brand Topic Maps." *SAS Global Forum Paper 3980-2016*.

Evangelopoulos, N., Shakeri, S., and Bennett, A. 2018. "Uncovering Tacit Knowledge in Twitter-Based Communities with Social Media Analytics." In J. Liebowitz (Ed.), *Data Analytics Applications* series, S. Hamwadeh & H.-C. Chang (Eds.), *Analytics and Knowledge Management* volume, 67–120. London: Taylor & Francis.

Hirschfeld, H. O. 1935. "A connection between correlation and contingency. " *Cambridge Philosophical Society Proceedings (Math. Proc.)*, 31:520–524.

Kabir, N. and Carayannis, E. 2013. Big data, tacit knowledge and organizational competitiveness. *Journal of Intelligence Studies in Business*, 3(2013):54–62.

Kimball, R., and Ross, M. 2002. *The data warehouse toolkit.* New York: John Wiley & Sons, 2nd Edition.

Nonaka, I. and Takeuchi, H. 1995. *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. New York: Oxford University Press.

Sharda, R., Delen, D., and Turban, E. 2018. *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*. Upper Saddle River, NJ: Pearson Education, 4th Edition.

RECOMMENDED READING

- *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS®*
- *SAS® Text Miner 14.2: Reference Help*
- *Getting Started with SAS® Enterprise Miner™ 14.1*
- *SAS/STAT® 13.1 User's Guide: The CORRESP Procedure*

CONTACT INFORMATION

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

Nick Evangelopoulos
University of North Texas
940-565-3056

Nick.Evangelopoulos@unt.edu
https://cob.unt.edu/users/nee0001