Paper 2070 - 2018

Fine Tuning Topical Content in Written Expression in Cloud-based Environments

Gurpreet Bawa, Sr. Manager, Accenture Solutions Pvt. Ltd.

Barry de Ville, Shashidhar Shenoy, Kaustav Pakira

ABSTRACT

Our recent work in the examination of variability in topic content among different social groups in social media demonstrates our ability to automatically detect different words and expressions that imply the same or similar meaning. The approach relies on first carrying out a vector-based topic identification that is computed across the entire conversational corpus in the social media collection. In the next step, various social groupings or sub-nets are identified. Social network membership and role (leader/follower) are identified so as to weight potential conversational influence. Finally, word and phrase lists are generated for each of the topic scores in each of the social groupings. The word and phrase lists are identified through rule induction through the use of either machine learning or specialized procedures such as the Boolean Rule generator in SAS® Text Analytics. By examining the overlap of common terms for topics among the various sub-nets, we can identify which descriptors apply across all social groups and which specialized, idiosyncratic, and idiomatic descriptors emerge in various sub-networks. This general approach is illustrated with an examination of various social groupings that are identified using the collected transcripts of all SAS Global Forum papers from the early 70s to the most recent.

INTRODUCTION

The evolution of cloud-based solutions places a premium on just-in-time calculated text products (as compared to other, older approaches that are based on relatively static taxonomies for example). In this respect the SAS Visual Text Analytics (SAS VTA) solution approach that is based on the real-time calculation of Text Topics has added appeal. The approach identified here is designed to add further value to SAS text topic products: we adapt recent text-related SAS Patents that have been specifically adapted to refine and expand on the traditional field-tested text topics that have been a part of the SAS text mining solution approaches since 2010. Methods are demonstrated that increase the precision of topic products and which can be used to semantically refine and disambiguate associated word terms. The notions of *Semantic Fields*, *polysemy* and hypernyms – approaches that are incorporated in the associated US Patent claims – are explained and illustrated.

References are included throughout, including references to the SAS Text Topic approach **(1).**

SAS text topics exploit a dimensional resolution approach adapted from classical factor analysis. One classic use of factor analysis is as a data reduction technique to identify multiple measurements that are aligned with a hypothesized latent dimension. The measurements that are aligned with a given dimension are taken to be various, overlapping indicators of that dimension. Similarly, in the factor analytic adaptation to text topics word token measurements that align with a given dimension are taken to be overlapping indicators of the “topic” that is constructed from the implied dimensional representation.

The approach, is illustrated in Figure 1 where we show the document collection, or “corpus”, as a term-document matrix.

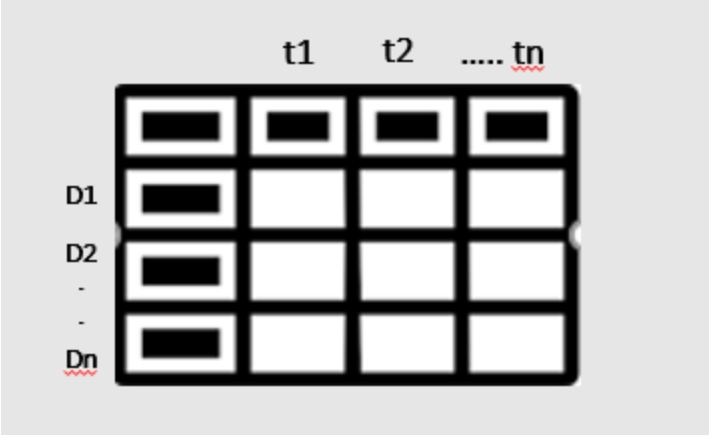


Figure 1: Illustration of Term X Document Corpus Representation

The term weight appearances that are inserted in the row X column cells are typically standardized across the entire collection: e.g. document term frequency / total term document frequency.
In this approach terms that align with the identified dimensions become term loadings on a Singular Value dimension and these are considered “topics”.

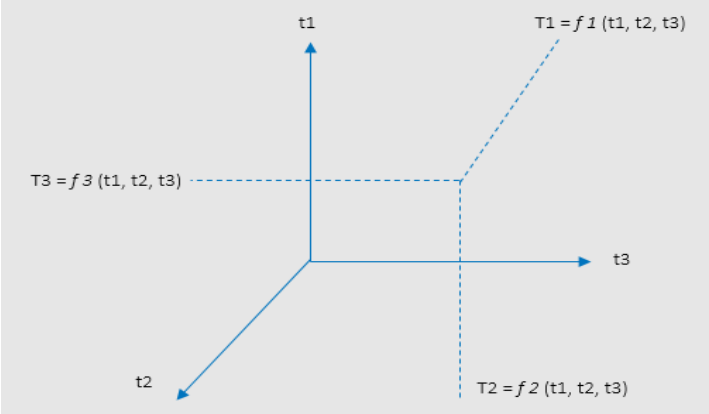


Figure 2: Illustration of Term (t', t'' ...) Associations with Dimensional Topic Products

The net effect is that each document in the corpus – collection has a potential “Topic” propensity for each of the identified topics (the Topic is given a positive occurrence in any given document when the topic factor score for that document exceeds a given threshold).

Document topic scoring pseudo code:

```
For Document1 through n
      For T1 through m
```

$$Topic_{nm} \leftarrow term_1 * c_1 + term_2 * c_2 + \ldots + term_n * c_n$$

Approach is well-adapted for cloud based solutions as Topics resolved in real time (e.g. less dependence on pre-calculated taxonomies)

NUMERIC-TO-SEMANTIC REPRESENTATION

Vector-based approaches can also be expressed as semantic products in line with a patent referenced in (2) (Cox and Zhao, 2014). The Boolean Rule generator turns calculated, vector-based factorial product into a semantic product. Figure 3 reproduces an example rule taken from the Cox and Zheng (2014) US Patent filing (3).

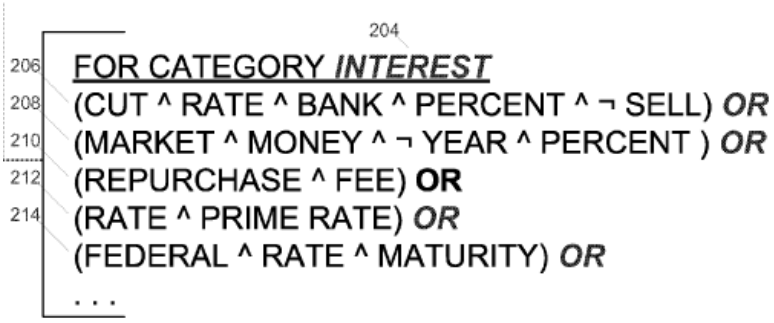


Figure 3: Example semantic (Boolean Rule) representation for numerical text product

SEMANTIC FIELDS, HYPERNYMS, POLYSEMY AND TOPIC ENUMERATION

While text topics capture the notion of descriptive terms that associate and vary with one another according to some underlying textual dimension we can also look at semantic variability in a document collection through the lens of a "semantic field". **In general, semantic fields describe a set of related words that describe the shared notion of a subject or concept.** The related words that are captured by the semantic field can be termed "hypernyms" in the sense that they share a common semantic property. In this sense hypernyms are (is-a) relations in the semantic field target. (4).

In this way pigeon, crow, eagle and seagull are all "is-a" descriptors – or hypernyms -- of bird or animal (5), The reverse of this expression leads to one example of the "polysemous" nature of word-terms; e.g. the word "bird" could be used to describe "pigeon", "crow" or "seagull" depending on the context of the term use. "is-a" relationships among hypernyms – in the context of terms and topics are therefore useful in determining specific sub-topic relationships among similar terms.

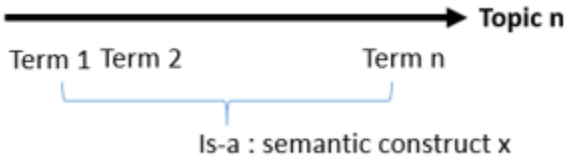


Figure 4: Term loadings on Topic Dimensions can be collected as Hypernyms "is-a" relations on associated semantic fields

In an earlier SAS Global Forum (SGF) paper (6) we extracted numerous networks of associated terms among papers that had been presented in SUGI or SGF presentations over a period of 1989 – 2012. An example of the semantic fields identified in this fashion overlaid on an illustrative Text Topic mapping is shown in Figure 6.

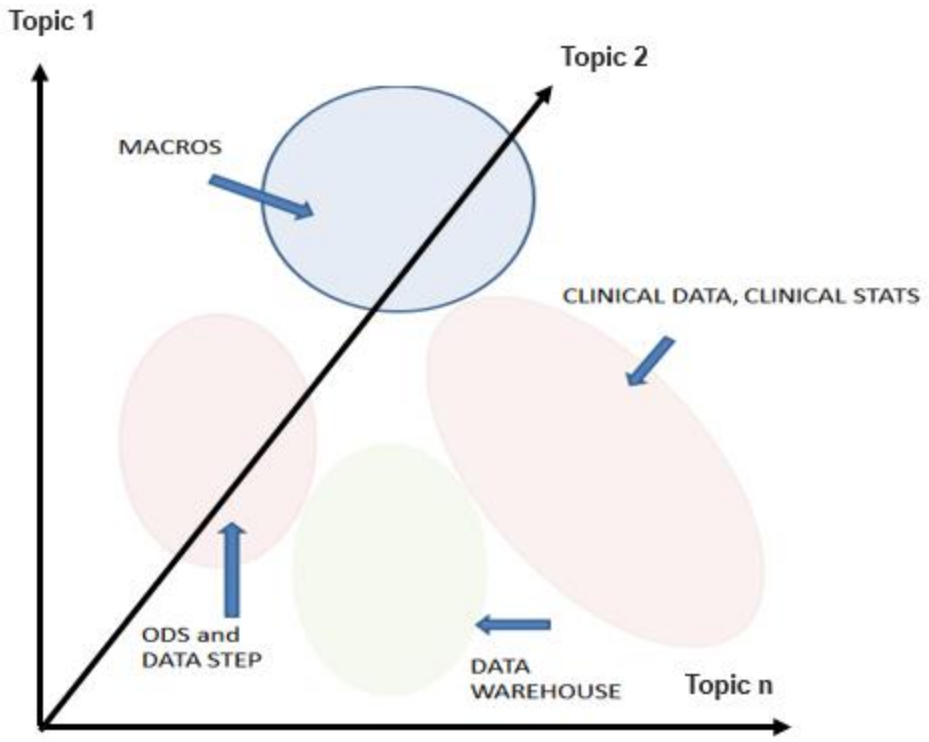


Figure 5: Semantic Fields Superimposed over Topic Representation

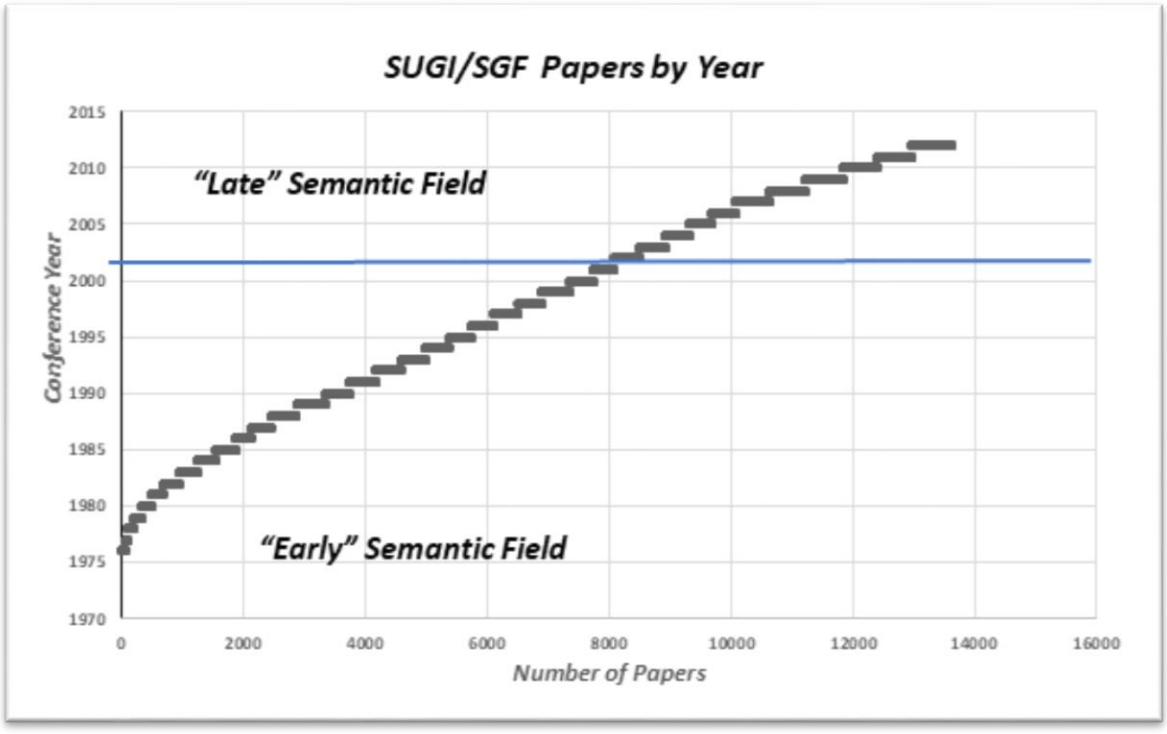
The goal of data reduction in the original factor analytic formulation means that terms that load on a given dimension take on the characteristics of the dimension according to the strength of the coefficient that accompanies the term x dimension association. The identification of terms with similar meanings – synonyms and hypernyms -- is secondary. If we adopt a definition of hypernyms as one or more terms that identify a given subject or concept then we can construct a semantic field – as an indicator of a given subject or concept – and superimpose over the topic resolution in order to identify hypernyms – terms that have a similar meaning in the context of the semantic field. This approach lends additional interpretive value to the topics.

HOW TO IDENTIFY HYPERNYMS WITHIN SEMANTIC FIELDS ASSOCIATED WITH A TOPIC

Semantic meaning and semantic fields are strongly related to social linguistic structure. Given this, when we construct social groupings over the collection of documents that are used as the corpus for the construction of Text Topics we give ourselves the opportunity to superimpose a semantic field structure over the text topic structure that is constructed: text topics are directly related to factor dimensions so do not directly incorporate notions of social groupings and associated semantic fields. This approach has been elaborated in other patents developed at SAS: (7) DeVille and Bawa, 2016 and (8) Cox et. al. 2016)

An Illustrative Example

Here we use the collection of SUGI/SAS Global Forum (SGF) data referenced earlier in (6) entitled “Two Technologies” paper SAS102-2013.



For purposes of our example we will create a "semantic field" segmentation that is based on paper issue data and will use papers from **1976-2002** as the **"Early semantic field"** segment and approximately equal number of papers from **2002 – 2012** as the **"Late semantic field"** segment.

Figure 6: Example Text Manipulations of "Two Technologies" SAS102-2013 DATA

Calculated topics across the entire range of papers then created two example semantic fields by dividing the papers into two – time defined -= sub-nets (or communities) based on papers from 1976 to 2001 – approximately one half – then papers from 2002 – 2012.

Notice: we work with topics that were computed across the entire corpus.

For purpose of this illustration we selected Text Topic 2. This topic has these associated term loadings:

- Data set
- SAS data
- Macro variable

- SAS Macro

With the entire data set we then extracted the associated Boolean rules

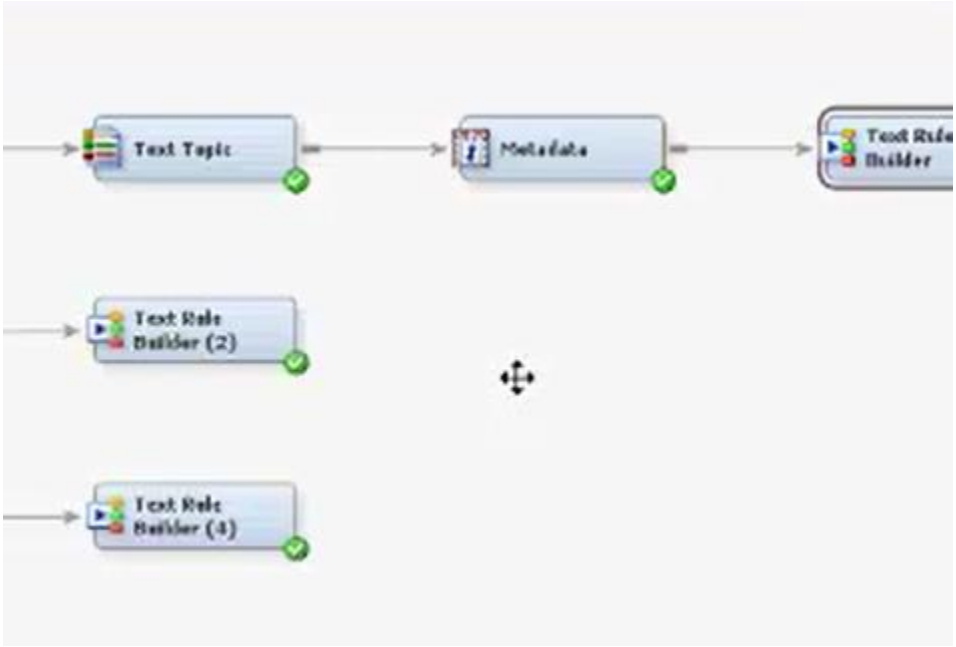


Figure 7: Using Boolean Rule Builder to find hypernym predictors for topic in 3 semantic fields

General Result

The extracted rules, illustrated in Figure 7 are enumerated below:

TEXT TOPIC NOT PRESENT (0) VERSUS PRESENT (1) … ASSOCIATED BOOLEAN RULE

Now extracted rules for Topic 2

TARGET_VAL	conj_id	conj_label
0	1	sas institute & ~sas data & ~data set
0	2	sas stat & ~data set
1	**3**	**data set & sas data**
1	**4**	**data set & macro variable**
1	**5**	**data set**
1	**6**	**integration studio**
1	**7**	**sas data & sas macro**
1	**8**	**sas data**

Figure 8: Boolean Predictors for Topic 2 -- Entire Dataset

N == 6659

Earlier data set – 1989 – 2002

Rules

TARGET_VAL	conj_id	conj_label
0	1	develop & ~set & ~sas data
0	2	statistical & ~data set
0	3	good & ~data set & ~sas data
0	4	company & ~data set & ~sas data
0	5	model & ~data set
0	6	service & ~set
0	7	technology & ~sas data
1	8	data set & sas data
1	9	data set & variable
1	10	sas data & set

Figure 9: Boolean Predictors for Topic 2 -- Early Data set (1976 - 2001)

N = 3320

Data from 2002 to 2012

TARGET_VAL	conj_id	conj_label
0	1	inc & ~sas data & ~data set
0	2	business & ~sas data
0	3	model & ~set
0	4	development & ~sas data
0	5	graphics
1	6	data set & sas data
1	7	data set & variable
1	8	integration studio
1	9	sas data & sas program
1	10	set & data set
1	11	sas data

Figure 10: Boolean Predictors for Topic 2 -- Later Data set (2002 - 2012)

N = 3339

SUMMARY RESULTS AND INTERPRETATION

A tabular display of the three sets of results is presented in Table 1. We can also align the results in line with the US Patent 9,582,761 filing approach which has been specifically adapted to display these kinds of results (9). This is shown in Figure 11.

Table 1: Comparison of Common and Different Terms (Hypernyms) But Sub-Net

Subset	Term Overlap	Disjointed Terms					
1976 - 2001	data set & sas data	data set & variable	sas data & set	~	~	~	~
2002 - 2012	data set & sas data	data set & variable	sas data	set & data set	integration studio	sas data & sas program	~
ALL	data set & sas data	~	sas data	data set	integration studio	sas data & sas macro	data set & macro variable

Interpretation

- The meaning of "data set & SAS Data is stable across the early, late and entire period
- "Data set and "Variable" -- distinct in early and late stages -- drop out when considering the entire period
- "SAS Data" and "Data set" evolve over time to have separate and distinct meanings
- The notion of "Integration Studio" emerged over time -- with the introduction of the product -- and the meaning persisted over time
- Over time the notion of "SAS data" evolved to support "SAS Programs". "SAS Macros" and the concept of a "Macro variable" emerged as a distinct kind of data.

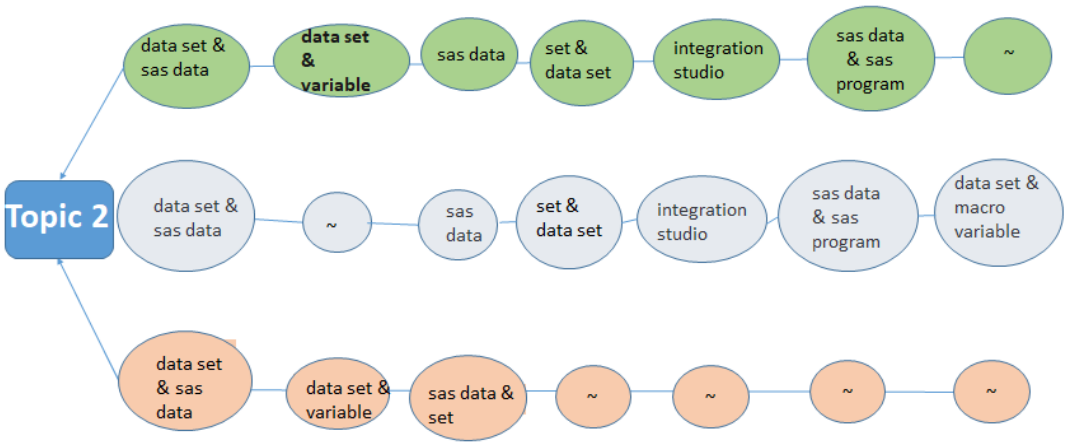


Figure 11: Distribution of Canonical Hypernyms using Topic as a target display in line with US Patent 9,582,761

CONCLUSION

This simple example demonstrates how an overlay of community-based semantic fields superimposed on Text Topics can lead to the extraction of unique terms – hypernyms – which capture unique meanings above and beyond the meanings extracted by the Text Topics themselves.

FUTURE RESEARCH

The example used here is purposefully simple. Other examples, could be constructed using the same data set and applying the semantic maps overlays identified in the Lavalley, DeVille, Bedford, Paper 102-2013 paper (6). A particularly useful overlay is shown in Figure 8. Here we identify semantic fields that are based on Author sub-communities. As noted in US Patent 9,317,594 filing (8), when the sub-nets are communicated with "author" as one of the components then influence metrics (e.g. centrality, authority) for the author can be calculated and this can be used to give a preferential weight to the terms that a particular author uses. One example use here would be to transform the authors "hypernyms" into preferred "synonyms". These synonyms could then be automatically re-mapped in the corpus and further text products could be calculated based on the automatic synonym re-mapping.

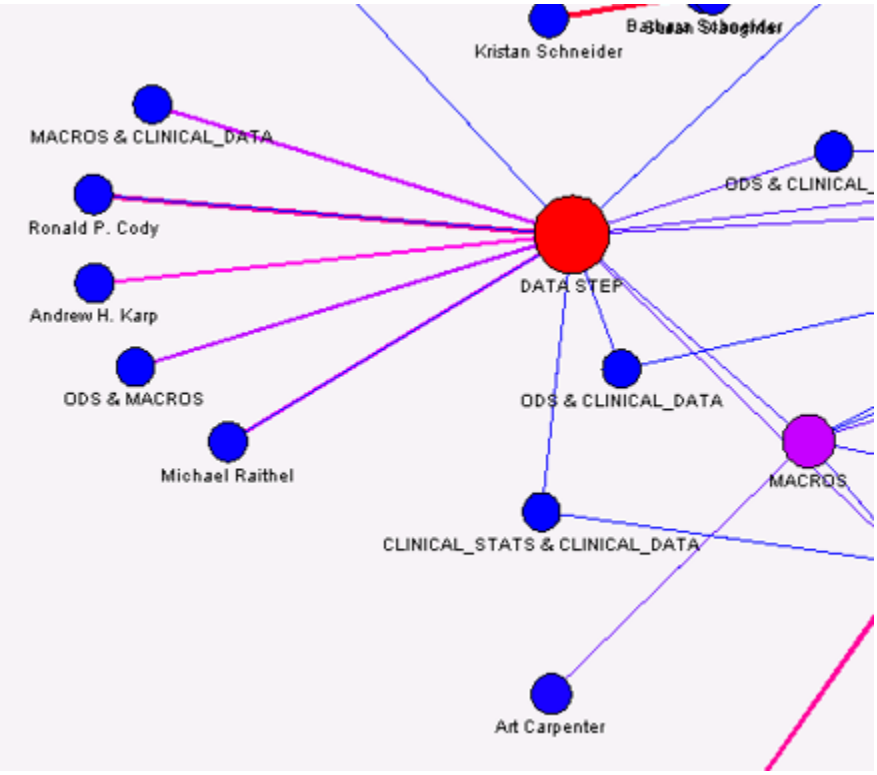


Figure 12: EXAMPLE Author-Topic SEMANTIC FIELDS in SUGI/SAS GLOBAL FORUM PAPERS

REFERENCES

1. SAS text topic node, e.g. Using the Text Topic Node :: Getting Started with SAS(R) Text Miner 12.1
   Links:
   https://blogs.sas.com/content/sasdummy/2010/05/27/a-topical-topic-how-text-mining-determines-topics/
   http://blogs.sas.com/content/text-mining/2010/04/16/the-whats-whys-and-wherefores-of-topic-discovery...
   http://blogs.sas.com/content/text-mining/2010/04/20/www-of-topic-management-part-2-what-is-a-topic-a...
   http://blogs.sas.com/content/text-mining/2010/05/12/part-3-understanding-topic-discovery-from-an-his...

2. United States Patent 8832015 B2 Fast binary rule extraction for large scale text data, 9 Sep 2014, James Allen Cox and Zheng Zhao

3. Orthogonal Transformation Figure 2 (Cox and Zhao, 2014)

4. Semantic Fields, Hypernyms, Polysemy and Topic Enumeration (4) (Snow, Jurafsky, Ng, 2004)

5. Hypernym descriptors for subnets : https://en.wikipedia.org/wiki/Hyponymy_and_hypernymy

6. 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 ...
http://support.sas.com/resources/papers/proceedings13/102-2013.pdf

7. United States Patent 9,317,594, April 19, 2016. Social community identification for automatic document classification, Inventors: De Ville; Barry, Bawa; Gurpreet

8. United States Patent 9,582,761, February 28, 2016. Generating and displaying canonical rule sets with dimensional targets, Inventors: Cox; James Allen, De Ville; Barry, Zhao; Zhen

CONTACT INFORMATION

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

Gurpreet Bawa
Accenture Solutions Pvt.Ltd
E-mail: gurpreet.singh.bawa@accenture.com

Shashidhar Shenoy
Accenture Solutions Pvt.Ltd
E-mail: shashidhar.n.shenoy@accenture.com

Kaustav Pakira
Accenture Solutions Pvt.Ltd
E-mail: kaustav.pakira@accenture.com