

Paper 109-2009

Discover and Drive Brand Activity in Social Networks

Barry de Ville, SAS Institute Inc., Cary, NC

ABSTRACT

This work report demonstrates how Web-based textual media can be mined and visualized to gain brand knowledge and customer leverage. Brands and key opinion leaders are identified, quantified, and visualized to answer such questions as:

- What is social media/consumer-generated content?
- What is social media marketing/monitoring?
- What is a niche community? Who are influencers? How are they identified?
- How do you identify brands?
- What are the appropriate reporting techniques and visualization approaches?

An end-to-end case study is presented using Web-based medical abstracts as illustrations of a range of social- and consumer-generated text sources that contain insights on brand adoption and use patterns.

Techniques—based on the interplay of SAS® Text Miner and SAS® Content Categorization powered by Teragram—are demonstrated as are social networks metrics calculations. Reports and exploratory social network visualizations through SAS/GRAPH® Network Visualization Workshop are presented.

Bottom-line benefits, extensions, and trends are reviewed.

INTRODUCTION

Text, brands and networks are now common-place terms that are often used to describe an emerging field of practice that scans web-based text content in search of brand use information. The Web scans attempt to identify trends, themes and tendencies among consumers that can indicate brand use, brand sentiment and attitudes toward brand manufacturers. Because these Web-mediated conversations are becoming increasingly common, they influence business and, because they influence business, no major manufacturer can afford to ignore these conversations. Consequently, more and more manufacturers and brand managers are making efforts to develop and improve upon the methodologies for conversation monitoring on the Web.

Text, Brands, and Networks is an end-to-end example of an emerging solution that uses a combination of SAS capabilities in text mining and analytics with social network and link analyses to uncover networks of brand references in Web-based media. This vision incorporates various methods to automate Web-mediated conversation monitoring. Automation is certainly required due to the increasing volume of content arising from these new forms of social interaction. Automation is applied to these new forms of interaction – often characterized as *social media* or *consumer-generated* media – in order to characterize and better understand these internet-mediated interactions between and among consumers.

The example illustrated here is modeled after work that was undertaken on behalf of a brand-management function within a major pharmaceutical manufacturer. The example demonstrates how Web-based textual media can be mined and visualized to gain brand knowledge for marketing interventions. Brands and key opinion leaders are identified, quantified and visualized to answer questions such as the following:

- How can a niche community of stakeholders be discovered and defined?
- Who are the influencers? How are they identified?
- How are brands identified?
- What are the appropriate reporting techniques and visualization approaches?

The example presented here uses the online database maintained by PubMed (www.pubmed.org). This database contains articles published by physicians on a variety of medical topics. Like blogs, emails and social networking site entries these articles use specialized terminology and use a variety of terms that are indicative of brands. PubMed maintains an Open Access Subset that includes articles from a number of publishers that have relaxed copyright restrictions to enable work of this kind. Because of the accessibility of this data and its utility to the manufacturer, it is

presented here as an example of the kinds of techniques and approaches that prove useful across a wide range of Web-based media in the monitoring of many brands and brand interactions.

THE SCENARIO

Historically, influences on physician-prescription practices have been tracked and monitored by means of custom physician surveys, either face-to-face, via telephone interview or, occasionally, by mail-in form. In traditional brand tracking methods, one or more surveys can be used. These surveys typically carry a high cost, require extensive follow up, and usually do not yield a high response rate (a 10% response rate is typical).

Automated methods, on the other hand, can plug into a variety of internet-mediated sources. Further, since all the sources are automatically scanned, the response rate is higher. Since the media is often freely available the cost can be relatively low – especially when compared to traditional methods.

In this example a web-based interface (PubMed) is used to access and retrieve information from a number of medical abstracts that are available in the Open Access Subset. The PubMed interface supports a File Transfer Protocol (FTP) agent that can be configured with keywords to pull back information. This information is subsequently analyzed for brand mentions and is presented and visualized using the techniques discussed here. In this example, we use an FTP agent to connect to the PubMed FTP-agent interface to retrieve information using the keywords NSAID and ASA. As we shall see, PubMed has enough built-in keyword mapping capability to recognize these keywords as proxies for a wide variety of anti-inflammatory agents, including popular brands that include these agents in both prescription and over-the-counter (OTC) form. These agents are often referred to by virtue of their chemical or physiological properties and effects. Much of the brand identification and term-mapping that takes place in this application is designed to map particular effects with particular brands. Disambiguation is also required: for example, “makes my head ache” might have a different meaning and interpretation than “I have a head ache” (there might be obvious differences in the intervention – chemical or otherwise – that are required).

THE PROCESS STEP-BY-STEP

The end-to-end process used in our example is shown in Figure 1:

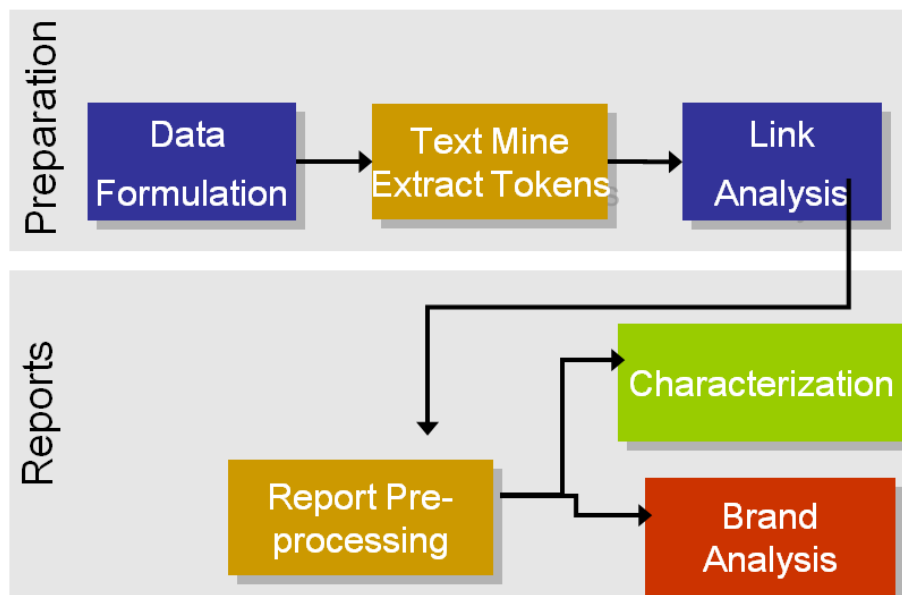


Figure 1: The Six Steps to Text, Brand and Network Visualization

As shown in Figure 1, there are six steps to the brand, text-mining, and network analysis process. Each of these steps has distinctive characteristics that produce a synergistic end product. Some of these distinct characteristics are highlighted in the following step descriptions.

DATA FORMULATION

The first step in the analysis is to gather the data. The PubMed Open Access Subset can be accessed via an FTP agent or manually through user access that uses a portal or Web interface. A significant advantage of the FTP-agent interface is that it returns data in extensible markup language (XML) form, as shown in Figure 2. XML is now a universally available file interchange format. Because the format enables the attachment of metadata descriptors to the fields of data that are being transferred, it is particularly useful in sending text information. *Metadata* is a form of

data labeling that informs users about the meaning and format of the fields contained in that data as well as the arrangement of these fields on information in the data set that is being transferred. As shown in Figure 2, XML is human-readable (rather than binary encoded, which is not easily readable).

```

</Journal>
<ArticleTitle>Cost-effectiveness of proton pump inhibitor cotherapy in patients taking
long-term, low-dose aspirin for secondary cardiovascular prevention.</ArticleTitle>
- <PageNumber>
<MedlinePage>1684-90; discussion 1691</MedlinePage>
</PageNumber>
- <Abstract>
<AbstractText>BACKGROUND: Patients with coronary heart disease (CHD) require
long-term therapy with low-dose aspirin (ASA). </AbstractText>
</Abstract>
<Affiliation>Department of Internal Medicine, University of Michigan Medical School,
3912 Taubman Center, Ann Arbor, MI 48109-0362, USA.
sdsaini@umich.edu</Affiliation>
- <AuthorList CompleteYN="Y">
- <Author ValidYN="Y">
<LastName>Saini</LastName>
<ForeName>Sameer D</ForeName>
<Initials>SD</Initials>
</Author>
- <Author ValidYN="Y">
<LastName>Schoenfeld</LastName>
<ForeName>Philip</ForeName>
<Initials>P</Initials>
</Author>

```

Figure 2: Example of XML Data Retrieved from the PubMed Data

As shown in Figure 2, all text fields are bracketed by XML tags that identify the field of information to which the text belongs. For example, the XML tag <Article Title> identifies the text element that contains the title of the article that is posted. Notice that the end of the title is indicated by the closing XML tag </ArticleTitle>.

For the sake of brevity, the full abstract text has been truncated in this illustration. Even this brief extract enables us to see the appearance of alternative terms – for example, “low-dose aspirin” and ASA – both of which are indications of a particular brand that is being searched for and analyzed.

Another characteristic of the analysis is also demonstrated in this figure. As shown in the XML data stream, this example contains multiple authors (here we see *Saini* and *Schoenfeld*). These two elements – terms that are indicative of a brand and authors that co-publish and co-reference one another – form the necessary components to animate the brand and network analysis that is shown here. Later, we see that authors who publish together can be used to identify self-referencing groups. Since authors publish with different authors on separate occasions, we can see how influential authors evolve. Such authors tend to publish a lot and tend to publish with a diverse community of other authors.

TEXT MINING/ANALYTICS

As shown in Figure 2, brand-indicating text can be extracted from such fields as AbstractText and ArticleTitle. Once the documents are placed in text format, this text can then be analyzed using a variety of techniques. In our case, we used a combination of both SAS Text Miner and SAS Content Categorization Server. One of the most useful features of SAS Text Miner is the ability to assign synonyms for various text terms that are identified in the text parsing. The synonym identification is accomplished through a user interface. This enables the rapid identification of many different text terms for a given brand or chemical substance. This process is illustrated in Figure 3.

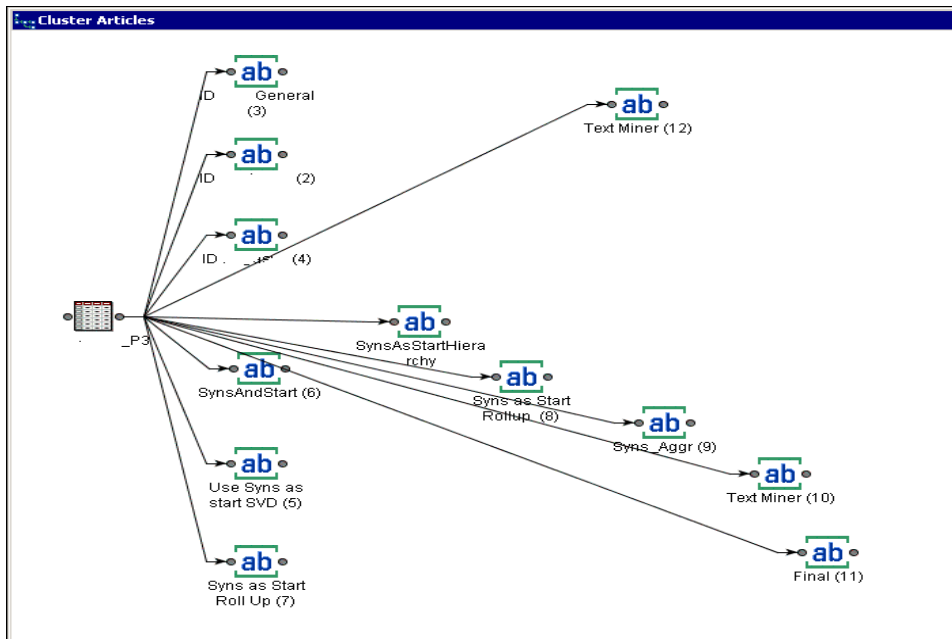


Figure 3: A Synonym Identification Diagram from SAS Text Miner

Initially the text terms are identified using a variety of Natural Language Processing (NLP) methods. The most common form of NLP tags the terms in text fragments by a part-of-speech identifier such as Noun, Verb, Adjective, Adverb, and so on. NLP also tags the document with entities. Entities that are identified include proper names, brands, addresses, and quantities. Even specialized descriptors, such as machine parts, are identified.

Other common NLP steps include stemming (creating term contractions), assigning synonyms, correcting spelling and such macro-document processes as identifying topics, sub-topics, and other parts of documents (such as major sections). The NLP capabilities of SAS Content Categorization are ideal for carrying out these types of tasks. These NLP capabilities are embedded in the SAS Text Miner interface, and they can also be used in a supplementary and complementary fashion *outside* the SAS Text Miner interface (as is done in this example). XML serves as the data-interchange mechanism when SAS Content Categorization and SAS Text Miner are used separately.

Figure 4 shows an example of the synonym table that is built up using the SAS Text Miner interface.

Term	Parent
acetylsalicylate	acetylsalicylate
acetylsalicylic	acetylsalicylate
acetylsalicylic acid	acetylsalicylate
acetylsalicylic acid sensitivity	acetylsalicylate
take acetylsalicylate	acetylsalicylate
taking acetylsalicylate	acetylsalicylate
asa	ASA
cox	cox-2
cox - 2	cox-2
cox-1/ cox-2	cox-2
cox-2	cox-2
cox-2 activity	cox-2
cox-2 inhibitor	cox-2
cox-2 inhibitors	cox-2
cox-2 pathway	cox-2
cox-2 pathways	cox-2
cox-2 therapy	cox-2
diclofenac	diclofenac
ibuprofen	ibuprofen
tiaprofenic acid	ibuprofen
lumiracoxib	lumiracoxib
naproxen	naproxen
naproxen nitroxybutylester	naproxen
naproxen sodium	naproxen
no-naproxen	naproxen
nsaid	nsaid
nsaid toxicity	nsaid

Figure 4: Example of Synonym Identification

Notice that synonym mapping accomplishes the major purpose of expressing various forms of terms into a common representation (sometimes referred to as the *canonical* representation). This significant function also incorporates spell checking and the mapping of chemical and physiological agents to associated brands. Canonical representations are also useful in term search applications (not discussed here) since search retrievals are greatly facilitated by common vocabulary, spellings, and meaning associations.

Figure 5 provides an example of how SAS Content Categorization identifies key terms in collections of text. In this figure, we can see that specialized terms are identified and are tagged with XML tags that clearly identify the terms. These tags are also useful in exchanging information between SAS Content Categorization and SAS Text Miner (and other SAS software, for that matter). Notice that the drug tags are also in the standard XML format, for example, `<ibuprofen>ibuprofen</ibuprofen>`.

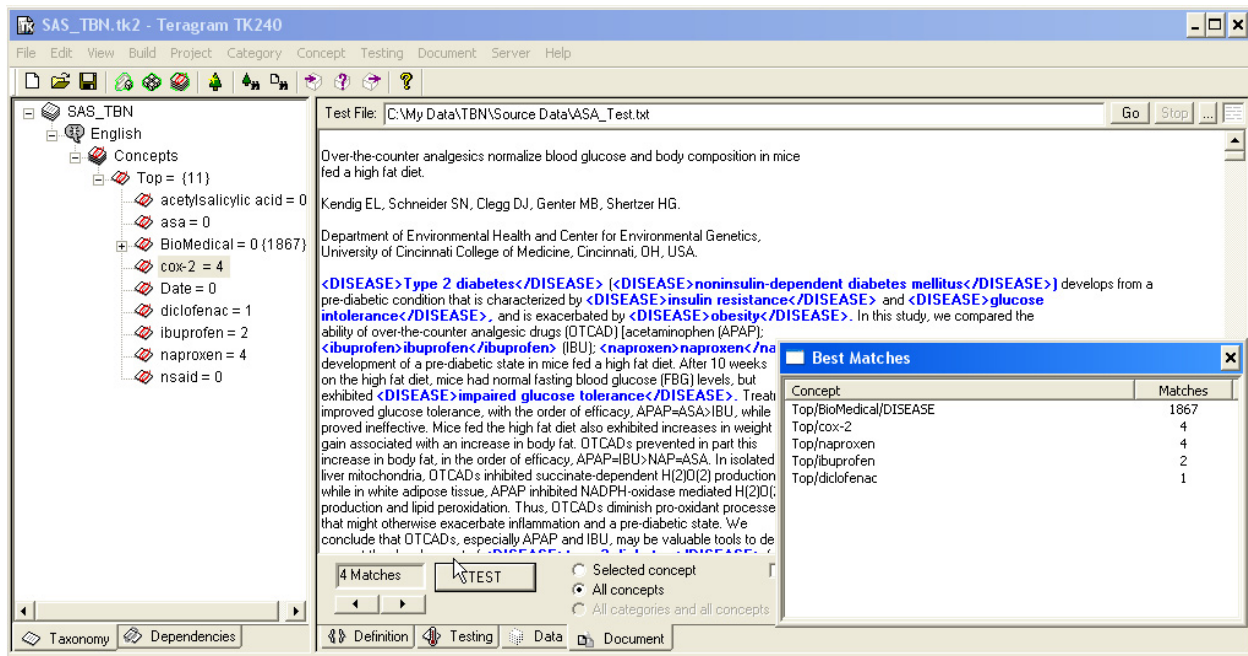


Figure 5: Example of Term Tagging in SAS Content Categorization Server

In our example, we also used the term-tagging capability of SAS Content Categorization Server as a mechanism to enable the user to change synonym mappings. Figure 6 illustrates a portion of the list of terms and synonyms that we had identified and which were inserted into an Excel workbook. This workbook was then reviewed by the product management team – before the analysis – to ensure that the terms were correctly identified and that the correct synonyms were applied. Because the product team had access to the complete synonym mapping scheme that was used in this project, they could reassign mappings based on their business experience. Figure 6 shows an example of this user-initiated re-mapping. For example, on line 4, the term ASA is substituted as a synonym for acetylsalicylic acid. The scoring abilities of SAS Content Categorization were subsequently used to apply the new synonyms to the data set that was used as input for the network analytics.

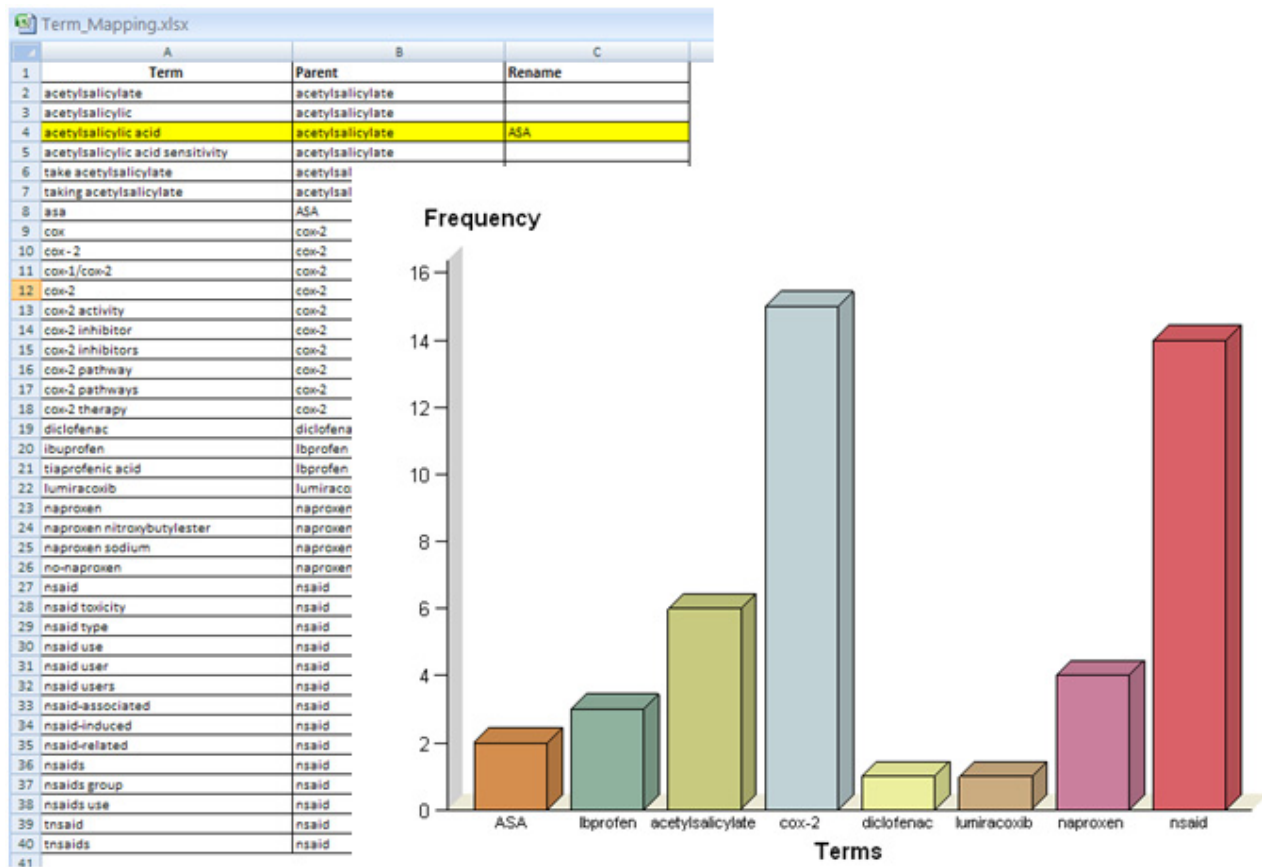


Figure 6: Example of Excel Workbook Used as a Host to Specific End-User Term Synonyms for the Analysis

LINK ANALYSIS/SOCIAL NETWORK ANALYSIS

Once the data is tagged with brand identifications, it can be transformed into a form that supports reporting and network analysis. The essential form becomes author – article pairings. The article contents contain any references to a brand that has been identified and mapped using the approaches discussed previously.

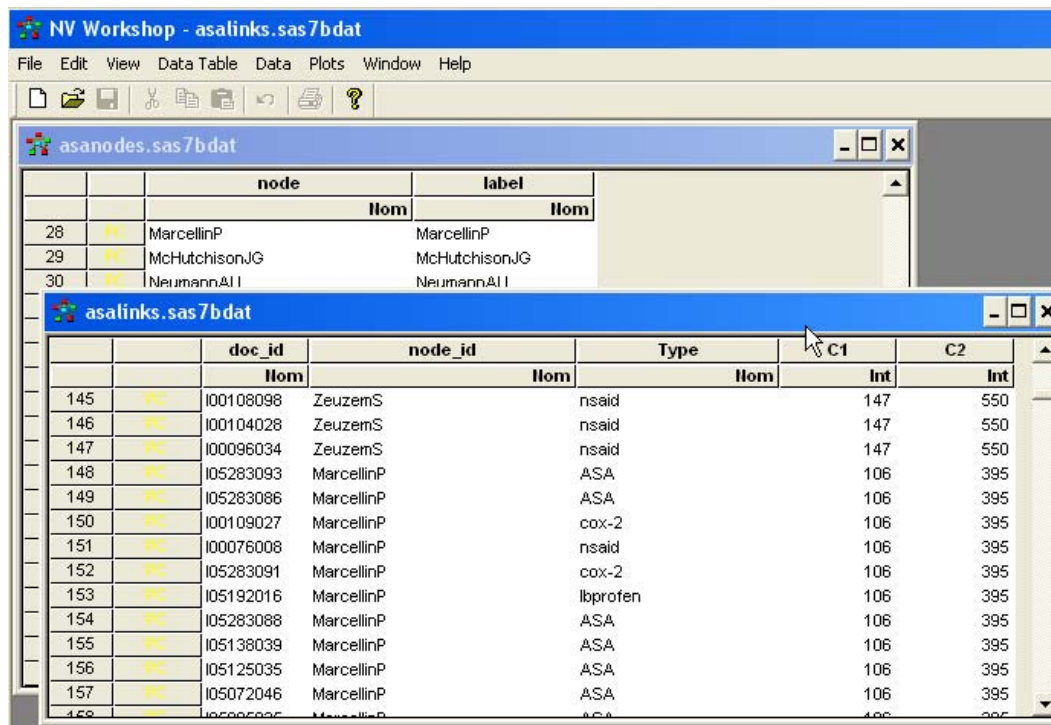
doc_id	node_id	Type	C1	C2
100138041	ZeuzemS	nsaid	147	550
100138041	BergT	nsaid	62	243
100138047	LawitzEJ	ibuprofen	35	133
100157003	PolS	cox-2	40	148
100157004	ZeuzemS	nsaid	147	550
100157004	SarrazinC	nsaid	48	185
100157008	WongJB	nsaid	36	131
100157013	ButiM	ibuprofen	34	131
100157035	PolS	nsaid	40	148
100157043	ZeuzemS	cox-2	147	550
100157043	BergT	cox-2	62	243
100157044	Diagom	nsaid	45	181
100157045	TrepoC	nsaid	53	202
100159040	SchalmSW	cox-2	42	161
100166021	LauGKK	nsaid	34	120
100173008	BergT	naproxen	62	243
100179002	EstebanR	naproxen	21	73
100179030	PoynardT	nsaid	54	195
100179030	WongJB	nsaid	36	131

Figure 7: Example of Data Reformatted for Analysis

Figure 7 shows a fragment of the analysis data set. Here we see that article 10013841 had two authors: Zeusem and Berg. For purposes of this analysis, the major brand that was discussed in the article was selected (the major brand that was discussed in the article usually outweighed minor brands mentioned by a considerable margin). In this

analysis, the goal was to track brand mentions. No attempt to attach a weight or sentiment ranking to the brand was attempted here.

The form of the data that is presented in Figure 7 was used for the production of all subsequent reports that are shown below. Network analysis usually requires the data to be placed in node – link form. Here, the authors are considered nodes and the articles that are published are treated as links between nodes.



The screenshot displays two SAS data tables. The top table, 'asanodes.sas7bdat', shows author nodes with columns for 'node' and 'label'. The bottom table, 'asalinks.sas7bdat', shows links between authors and articles, with columns for 'doc_id', 'node_id', 'Type', 'C1', and 'C2'.

node		label	
	llom		llom
28	MarcellinP	MarcellinP	
29	McHutchisonJG	McHutchisonJG	
30	NeumannAI I	NeumannAI I	

doc_id		node_id		Type	C1	C2
	llom		llom	llom	Int	Int
145	I00108098	ZeuzemS	nsaid		147	550
146	I00104028	ZeuzemS	nsaid		147	550
147	I00096034	ZeuzemS	nsaid		147	550
148	I05283093	MarcellinP	ASA		106	395
149	I05283086	MarcellinP	ASA		106	395
150	I00109027	MarcellinP	cox-2		106	395
151	I00076008	MarcellinP	nsaid		106	395
152	I05283091	MarcellinP	cox-2		106	395
153	I05192016	MarcellinP	lbprofen		106	395
154	I05283088	MarcellinP	ASA		106	395
155	I05138039	MarcellinP	ASA		106	395
156	I05125035	MarcellinP	ASA		106	395
157	I05072046	MarcellinP	ASA		106	395

Figure 8: Example of the Node and Link Data Representation Used in Link Analysis

Figure 8 shows an excerpt from the node and link data representation. The top of the figure shows the author nodes and node labels (contained in the asanodes data set). The bottom figure shows the corresponding links. These links enumerate the connections between authors and articles. The **Type** field indicates the brand term that has been identified as the major brand that is contained in the article.

The **C1** and **C2** fields are the indicators of first degree and second degree network influence, respectively. The first degree metric calculates the number of direct links that exist between authors and articles. Thus, the value of C1 is the number of articles that an author has published. The second degree metric is a calculation of *both* the number of articles that the author has directly published in *addition to* the number of articles that are published by the authors *with whom* the author publishes. Thus, the value of C2 is a combined measure of direct publications and indirect (by one link) publications. Since only co-authors are considered in the indirect measure, this is taken as a second degree metric.

There are many ways to calculate C2. In this case C2 was calculated using the link analysis macros originally released in the link analysis facility of SAS® Enterprise Miner 4.3™.

REPORT PRE-PROCESSING

A unique feature of this example included the ability to identify the top 20% of authors. To do this, quintiles are produced whereby the data is grouped into fifths, with authors ranked from top to bottom. Rankings, and associated quintiles, included C1 and C2 metrics as well as brand mentions. The intention is to harness the Pareto Principle in subsequent reports so that the top 20% of the authors are used to identify the most frequent 80% of brand mentions.

CHARACTERIZATION

Figure 9 shows an example of the kinds of reports that are produced to characterize the collection.

Author	Number of Publications	Span
Zeuzem, S.	147	550
Marcellin, P.	105	395
Ferenci, P.	88	324
McHutchison, J. G.	83	303
Shiffman, M.	84	299
Berg, T.	62	243
Kumada, H.	60	238
Trepo, C	53	202

Figure 9: Typical Profile Report Used to Characterize the Data

Here we see the enormous influence that is demonstrated by the top few authors in our example. The major author has almost twice as many publications as any of the other authors in the top 20% of the articles in the collection. This author's calculated C2 metric – labeled **Span** in the report – is over three times as great as the C1 metric (**Number of Publications**). This gives this author a significant influence (and also indicates that the authors that he co-authors with have a significant influence – and associated C1 metric – in their own right).

Later, in the section on network visualization, we will see that Marcellin, the second most-influential single author in this collection, interacts with another separate group of co-authors that he publishes with. So Zeuzem and Marcellin have built up their respective networks and influence metrics in autonomous fashions. Later, we identify two *bridge* authors who do, in fact, co-publish with Zeuzem and Marcellin. These “bridge authors have a disproportionate influence by virtue of the fact that they are scarce connectors between these otherwise autonomous circles of influence.

BRAND ANALYSIS

This report shows the coverage of company brand (**Our Brand**) versus competitor brand (**Their Brand**). In this case, we see that the top author favors the company brand. Five rows down, we see an author who favors the competitive brand (3 company brand mentions versus 12 competitive brand mentions).

Author	General	Our Brand	Their Brand	Total
Zeuzem, S.	49	41	14	147
Marcellin, P.	31	42	11	105
Ferenci, P.	36	23	6	88
Shiffman, M. L.	24	27	5	84
McHutchison, J. G.	36	3	12	83
Berg, T.	25	11	6	62

Figure 10: Example Brand Analysis Report

Another important analysis relies on the identification of the top 20% of the authors that account for up to 80% of the brand mentions in the collection of articles contained in the study. This was identified as the *Our Brand Area* in the Figure 11. The authors were further classified according to all brands mentioned, regardless of whether the company or competitive brand was referenced. By super-imposing the top members in both quintiles, we can identify a subset of articles that contain most of the brand mentions in this collection of over 13,000 articles.

		Category General					
		Q5	Q4	Q3	Q2	Q1	Total
Our Brand Area	Q5	276	68	0	0	0	344
	Q4	0	695	922	0	0	2267
	Q3	0	0	565	3122	3048	6733
	Q2	0	0	0	0	2267	1617
	Q1	0	0	0	0	2049	344
Total		276	763	1487	3122	7362	13010

Figure 11: Quintile Analysis

Figure 11 illustrates a sweet spot of 1039 authors who have leading expertise both in specific disease and in the Our Brand area as whole. This “sweet spot” represents a “high leverage” segment of the brand-influencing population.

NETWORK VISUALIZATION

As shown in Figure 12, there are many ways to visualize a network. In this figure, we can see a fully connected raw network in the upper-right quadrant. This subview shows all the connections that exist between authors and articles. Because there are so many connections the visualization is not very useful. The sub-view in the lower-right quadrant of Figure 12 shows a portion of the network where only nodes with high C2 (**Span**) are highlighted. This subview demonstrates how one particularly influential node can be identified.

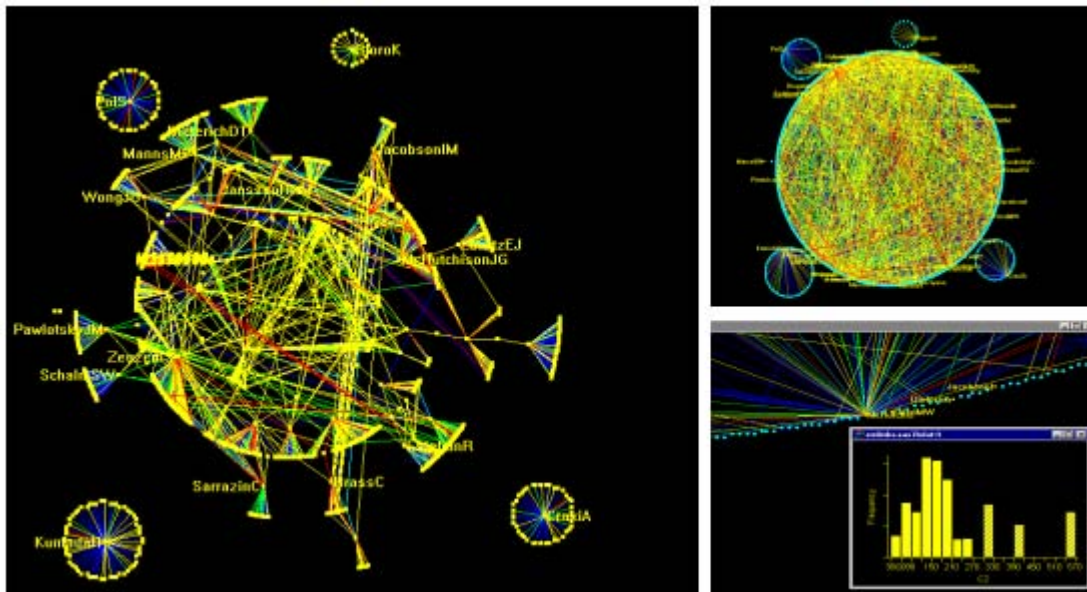


Figure 12: Example Network Visualizations

The view on the left side of Figure 12 is perhaps the most useful of all the views presented. Here we see displayed only those nodes with high C2 (**Span**) metrics. Because the links that connect the nodes are color-coded (based on the brand that is mentioned in the article), we can quickly identify high-influence authors that discuss ASA topics (shown in yellow) versus those that discuss NSAIDs (shown in red). Other brand mentions – for example, cox-2 inhibitors – are shown in other colors (such as blue and green).

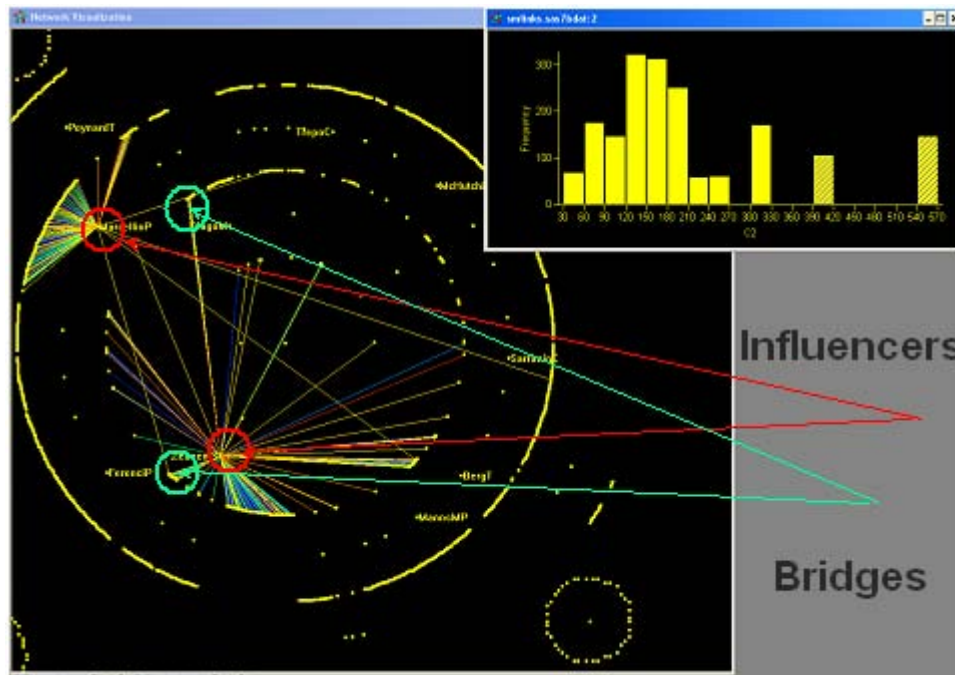


Figure 13: High Influence and High Connectivity Nodes

Figure 13 contains a further refinement on the left hand side that was presented in Figure 12. The legend view in the upper-right quadrant illustrates that only those nodes with very high C2 (**Span**) metrics – 400 and above – are displayed. This visualization provides a dramatic demonstration of how visualization can yield insights that would not be readily apparent using other methods. In this case we see that the two major authors in the study – circled in red – appear as separate *islands of influence*. Yet two authors – circled in green – serve as bridges between these otherwise separate influencers. Because these bridges link these two influence leaders, they take on an importance that is considerably greater than the influence of their individual publications alone. This is a clear illustration of the importance of connections in networks of interaction. It also shows how high-status nodes can be identified. Such nodes characterize authors who write articles with one another, but not with others in their field. They have high status and high *betweenness* scores by virtue of the influence and connectedness of the authors with whom they publish.

Figure 14 shows a network visualization complement to the type of report that was originally presented in Figure 10. In this figure, we can see the brand mention tendencies of the various authors. As shown in the legend in the upper-right quadrant of the figure, mentions of ASA (blue) and Naproxen (yellow) are highlighted in this visualization. This illustrates how authors with various brand-mention tendencies can be identified. It also shows whether authors have a predominant brand-mention preference or whether they mention a variety of brands in the articles that they publish.

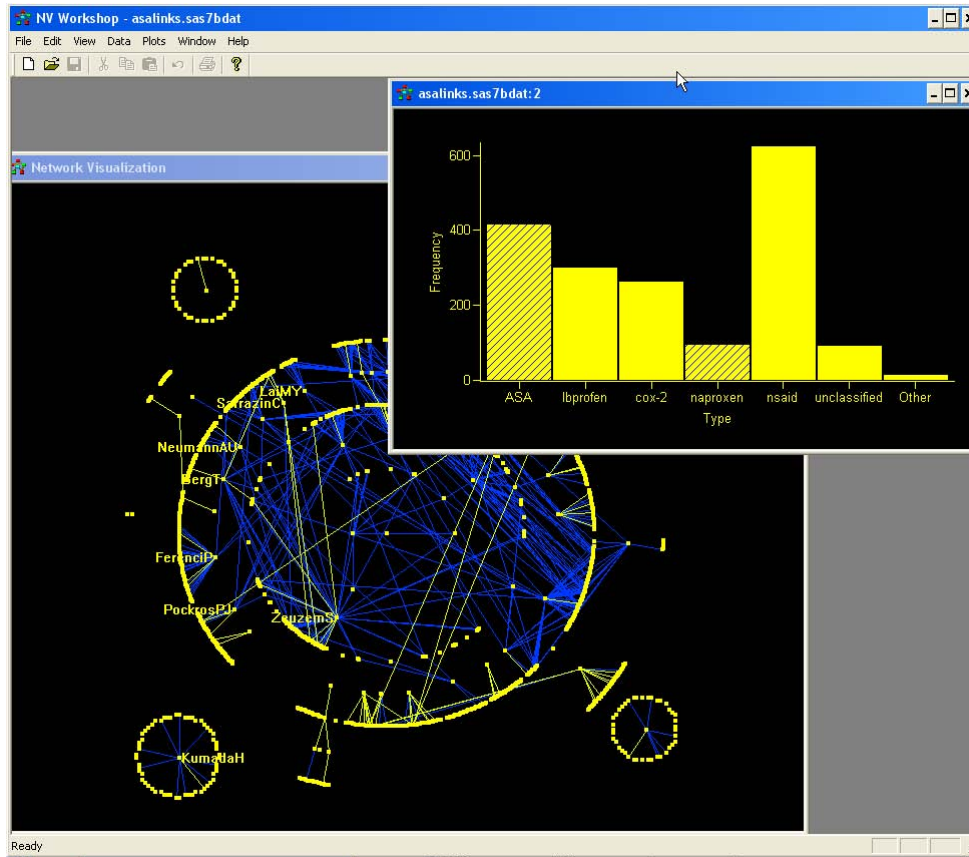


Figure 14: High-Influence Authors by Brand Mention

CONCLUSION

The scenario presented here demonstrates an end-to-end, worked-out example of how text from the Web can be captured, tagged, and analyzed to extract brand conversations. The important role that is played by identifying terms and term combinations that can be used to identify a brand is highlighted. The importance of synonym mapping and the interplay and synergy between SAS Text Miner and SAS Content Categorization server is illustrated. Further information about these techniques and the various forms of term taxonomies and synonyms that are available to do this work is available from the author. (See [Contact Information](#).)

Once the text terms are extracted, then a document with authors and terms is assembled. This document forms the main data set for the production of a variety of reports that can be used to identify brand trends and influencers. This data can be further refined into a node-and-link data set to form the basis of network visualizations. The network visualizations demonstrate how influencers that are identified in report format can be similarly identified through means of visualizations. And just as reports can show the relative proportions of brand references among authors, so, too, can these relative brand references be identified through network visualizations.

The special properties of network visualizations were also demonstrated. In this case, the importance of betweenness indicators in the identification of bridges was shown. Bridges have high status by virtue of the connectivity that they inherit from the influential nodes to which they are connected. They can play a significant role in the mediation of the conversation between separate – and otherwise autonomous – groups of influencers.

The information that is identified here can serve as an important guide to following brand conversations, for tracking customer sentiment and for constructing responsive interactions with customers that are based on knowledge gained through text, brand, and network analytics.

ACKNOWLEDGMENTS

I would like to acknowledge the significant contributions of Glenn Abrahamsen and Holly Fedeyko in the early development and review of this work. Patrick Homer, Steven Law, and Phil DiMassimo at SAS Institute Inc. have provided significant commentary and encouragement in the evolution of this work.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. For more information and detail on any of this content please contact the author at:

Barry de Ville
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
Work Phone: 919-677-8000
Fax: 919-677-4444
E-mail: barry.deville@sas.com
Web: www.support.sas.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.