Paper 076-30

Combining SAS® Text Miner with the Association Node in SAS® Enterprise Miner(tm) to Investigate Inventory Data

Patricia B. Cerrito, University of Louisville, Louisville, KY

ABSTRACT

Association rules are developed particularly using records of purchases to increase sales. However, when the inventory is large, the number of items can overwhelm relatively modest servers, requiring considerable investments in hardware. The number of items also suggests that the strength of any association will be low. Inventory items that are similar generally have similar sounding names. While the inventory codes are sometimes developed to indicate broad classes of items, it is not always possible to do so. For example, cereal in a grocery store has only a handful of categories that are too broad for association purposes. The names of items, however, can be used with the Text Miner Node prior to use with the Association Node. The combination of analyses will be used with an almost unlimited inventory of items that are used to treat patients in a hospital ER. Over a 3-month period there were almost 500 different items used for approximately 3000 patients, and almost 30,000 items charged to patients. The number of items and charges overwhelmed a Xeon dual-processor with 4 gig of RAM. The combination of text analysis and association rules can give very powerful results, in a manageable combination of hardware.

INTRODUCTION

An association rule is one of the most powerful data mining techniques. The goal of the method is to find interesting associations and correlation relationships amongst large sets of data items where the presence of one set of items in a transaction implies the presence of other items. For example, in a grocery store, it is important to determine whether customers who purchase beef also purchase marinades for the beef, and whether these items should be in close proximity to each other. Association rules examine the strength of the purchase combinations. Given a set of transactions where each transaction is a set of items, an association rule is an implication of the form $X \rightarrow Y$ where X is the set of antecedent items and Y is the consequent item.

One of the biggest problems with using the Association Node is the number of items under consideration. Many different categories take incredible amounts of computing time and memory. The task can easily exceed the grasp of desktop computers, or even of servers. When there are too many categories in the target variable, it is better to combine them in such a way as to reduce their number to a more manageable few. One of the ways to do this is to use the Text Miner Node. To demonstrate how the Association Node can be enhanced by the Text Miner Node, a database containing almost 30,000 charges for 3000 patients treated in a hospital ER were examined. In this example, the choices are made by physicians in the ER. The combination extracts meaningful intelligence from the database.

ENTERPRISE MINER

In addition to the antecedent X and the consequent Y, an association rule has two numbers that express the degree of uncertainty about the rule. In association analysis the antecedent and consequent are sets of items called item sets that are disjoint ($X \cap Y = \emptyset$). The first number is called the support for the rule. It is the number of times that the combination appears. The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule. The other number is known as the confidence of the rule. Confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent to the number of transactions that include all items in the antecedent. The association node requires one identification input and one target value (Figure 1). The targets give the inventory items purchases or used by the individuals contained within the ID field. The screens (Figures 2,3) give the Association Node defaults.

Figure 1. Association Node in Enterprise Miner

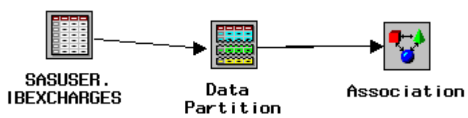


Figure 2. Basic Defaults for Association Node

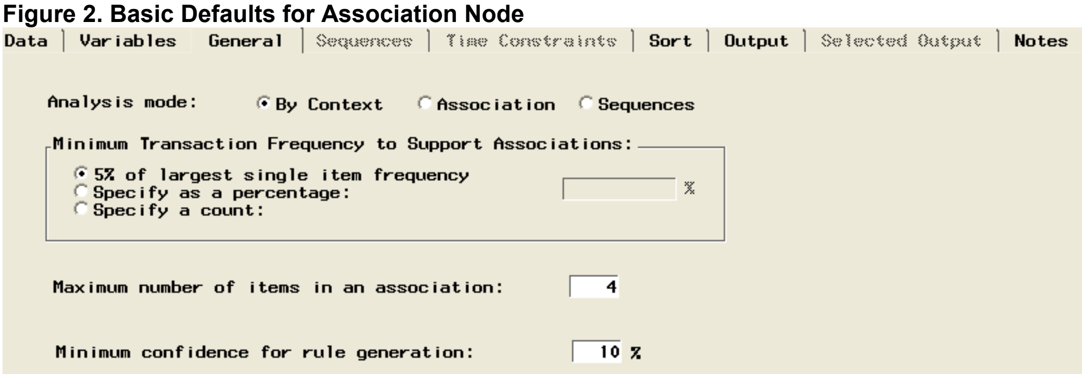
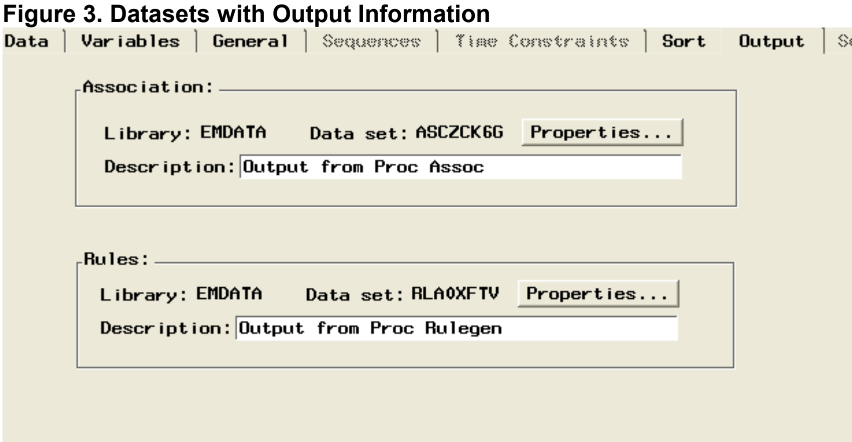


Figure 3. Datasets with Output Information



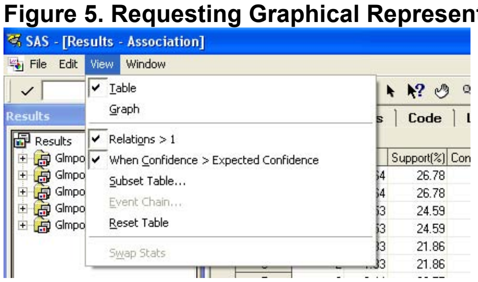
Basic results are given in Figure 4.

Figure 4. Basic Results of Association Node

	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
1	2	1.09	70.00	87.50	7.00	X-Ray ==> Diabetes & Asthm
2	2	1.09	70.00	87.50	7.00	Diabetes & Asthm ==> X-Ray
3	2	1.11	60.00	100.00	6.00	Heart Monitoring ==> Antibiotic
4	2	1.11	60.00	66.67	6.00	Antibiotic ==> Heart Monitoring
5	2	1.04	50.00	83.33	5.00	Heart Monitoring ==> Diabetes & Asthm
6	2	1.04	50.00	62.50	5.00	Diabetes & Asthm ==> Heart Monitoring
7	3	1.09	70.00	87.50	7.00	X-Ray ==> Urinalysis & Diabetes & Asthm
8	3	1.09	70.00	87.50	7.00	Diabetes & Asthm ==> X-Ray & Urinalysis
9	3	1.09	70.00	87.50	7.00	X-Ray & Urinalysis ==> Diabetes & Asthm
10	3	1.09	70.00	87.50	7.00	Urinalysis & Diabetes & Asthm ==> X-Ray
11	3	1.09	70.00	87.50	7.00	X-Ray ==> Lab Tests & Diabetes & Asthm
12	3	1.09	70.00	87.50	7.00	Diabetes & Asthm ==> X-Ray & Lab Tests
13	3	1.09	70.00	87.50	7.00	X-Ray & Lab Tests ==> Diabetes & Asthm
14	3	1.09	70.00	87.50	7.00	Lab Tests & Diabetes & Asthm ==> X-Ray
15	3	1.09	70.00	87.50	7.00	X-Ray ==> IV Charges & Diabetes & Asthm
16	3	1.09	70.00	87.50	7.00	Diabetes & Asthm ==> X-Ray & IV Charges
17	3	1.09	70.00	87.50	7.00	X-Ray & IV Charges ==> Diabetes & Asthm

Values with high support and high confidence represent strong associations. The associations are given in terms of the number of items in the association, and in decreasing order of support. With so many rules, it is helpful to use graphical representation (Figure 5).

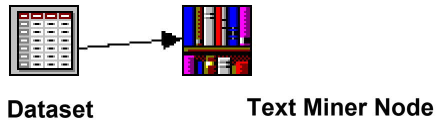
Figure 5. Requesting Graphical Representation



The horizontal axis is the predicted target value versus the actual target value. The vertical axis is either the frequency count or the percentage. The confidence is given by the shape on the graph; the support by the color.

The initial icons for Text Miner are given in Figure 6. These nodes can be integrated into Enterprise Miner provided that Text Miner is available.

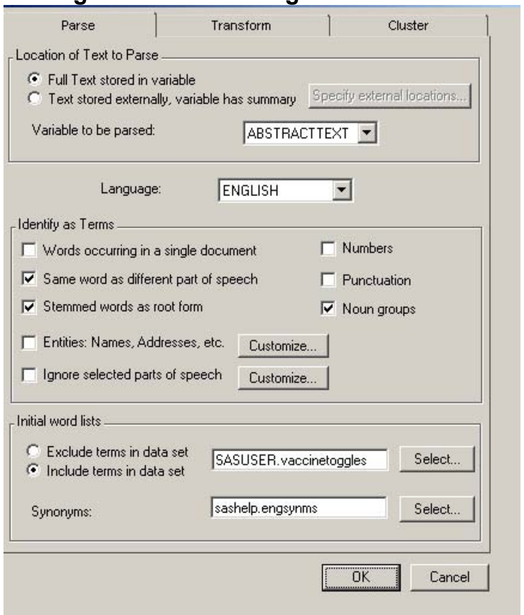
Figure 6. Text Miner Node



The Text Miner Node has three settings screens to examine. The first screen is given in Figure 7

.

Figure 7. First Settings Screen



There is an option to choose if the text is stored in a SAS dataset, or if there is a variable in the dataset that points to the location of the document. This second option is available to reduce the required storage size for a SAS dataset. In the second option, there is no limit on the size of each document; for the first option the size is restricted to 10 pages.

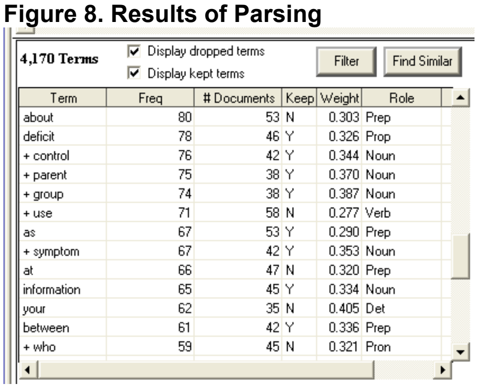
The first default is to exclude consideration of words that only occur in one document since those words cannot be used to group documents together. Unchecking the box will also create a much longer wordlist. A second default is to consider a word to be different if it is used as a different part of speech.

It is often advantageous to uncheck this box since Text Miner sometimes has difficulty with grammar. It should be unchecked when using Text Miner for inventory lists. In addition, it is possible to ignore specific parts of speech (such as conjunctions). Words with the same stem should be considered the same word. Therefore, the third box should always be checked. Numbers and punctuation are not ordinarily used to cluster text documents as well. However, inventory codes can be examined using text miner. If that is the case, then the numbers box should be checked. It is suggested that the user experiment with the defaults to observe the impact on the final outcomes.

A standard “stoplist” dataset will remove common words such as “and” and “the” from consideration. Users can add words to the stoplist as needed, or create their own lists. Similarly, a “startlist” can be defined. In this case, only the specific words contained within the startlist will be used. This option is useful to flag documents containing particular words or phrases. It differs from a more typical keyword search in that several words and phrases can be searched simultaneously.

It is possible to restrict attention to some specific terms by listing them in a dataset. Text Miner will only list terms from the specified dataset. The purpose of this step is to ‘parse’ the documents. Text parsing is a very technical process that is used to reduce the size of the documents to a manageable number. It also means that the software attempts to use grammar context to identify a specific part of speech for each term used. Modifiers are often connected to nouns to define ‘noun groups’ (Figure 8).

Figure 8. Results of Parsing



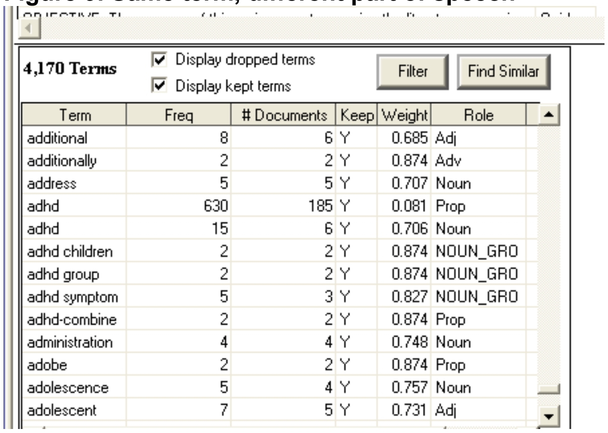
4,170 Terms — Display dropped terms — Display kept terms — Filter — Find Similar

Term	Freq	# Documents	Keep	Weight	Role
about	80	53	N	0.303	Prep
deficit	78	46	Y	0.326	Prop
+ control	76	42	Y	0.344	Noun
+ parent	75	38	Y	0.370	Noun
+ group	74	38	Y	0.387	Noun
+ use	71	58	N	0.277	Verb
as	67	53	Y	0.290	Prep
+ symptom	67	42	Y	0.353	Noun
at	66	47	N	0.320	Prep
information	65	45	Y	0.334	Noun
your	62	35	N	0.405	Det
between	61	42	Y	0.336	Prep
+ who	59	45	N	0.321	Pron

The + signs indicate that there is more than one word connected to the phrase. Clicking on the ‘Term’ box will put the words in alphabetical order.

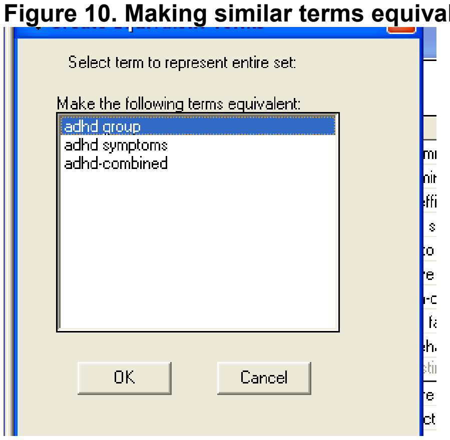
Notice that some of the terms have a ‘Y’ or an ‘N’. Any value with an ‘N’ is contained within the ‘stoplist’ file and is not used in the analysis. By unchecking the box, ‘Display dropped terms’, all values with an ‘N’ are removed from the window; unchecking ‘Display kept terms’ removes all words with a ‘Y’. Consider Figure 9.

Figure 9. Same term, different part of speech



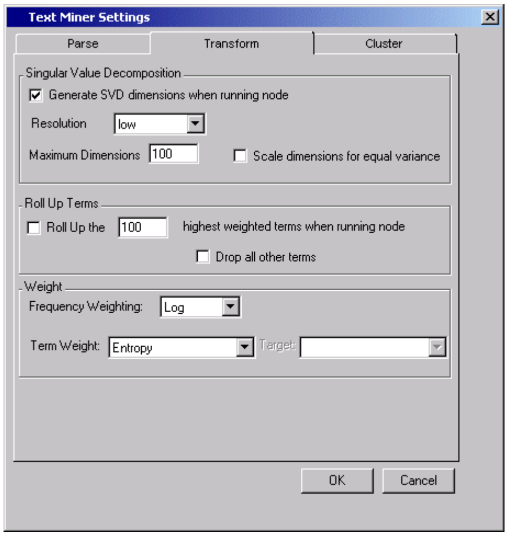
The term, 'ADHD' is both a proposition and a noun, as well as contained within a noun group. Text Miner will not allow these terms to be made equivalent across different parts of speech will result in an error. The noun groups can be made equipvalent because they have the same part of speech (Figure 10).

Figure 10. Making similar terms equivalent



This particular example originally generated 4,170 different terms. By unchecking the default 'Same word as different part of speech', reduces the number of terms to 3,914. Now it is possible to combine all references to 'ADHD' since parts of speech are no longer considered different.

Figure 11. Second Settings Screen



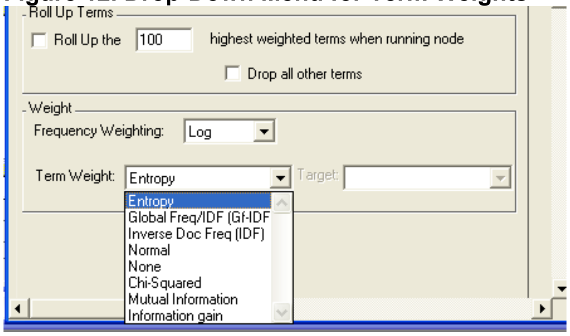
The second screen allows for the user to determine the method of reducing the wordlist matrix to a manageable size. The default is to use singular value decomposition. There are also several possible methods to weight the value of each term in the documents.

To investigate how these weights and methods impact outcomes, it is best to use one dataset and change the settings to see how the results differ.

The number of dimensions defaults to 100. However, that number can be decreased for a smaller number of documents, and increased for a large number (although the time factor will increase considerably.

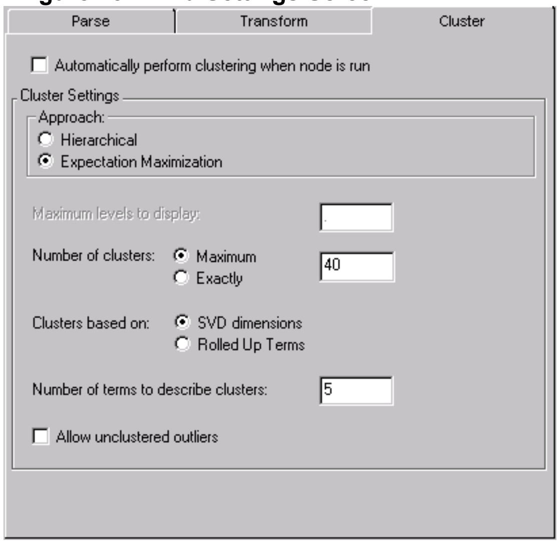
Singular Value Composition defines a matrix of words by documents. The maximum dimensions (by default 100) box limits the size of this matrix. However, the larger the matrix, the more time-consuming this process. The roll-up terms limits the wordlist to the top (100) highest weighted terms. It is suggested that the user modify the dimensions somewhat to determine the impact on the final outcome of Text Miner. Once the singular value decomposition is run, briefly, a status screen pops up, indicating that the singular value decomposition is being performed. The user can close this screen since the process will continue to run. A drop down menu will allow the user to change the weights (Figure 12). Common words such as "and" that are not specifically listed in the "stoplist" dataset should not be given a large weight since almost all documents contain many "and" words and they contribute very little to grouping documents.

Figure 12. Drop-Down Menu for Term Weights



Entropy is the default weighting. Terms that appear more frequently will be weighted lower compared to terms that appear less frequently. This weighting is somewhat different Inverse Doc Freq where documents that appear in as few as 2 documents are given the highest weights.

Figure 13. Third Settings Screen



Unless the box is checked, clustering is not automatically performed. However, once Text Miner completes the parsing and transformation steps, the user can request that the clustering be performed by using the settings value in the Tools pull-down menu in the results display.

The user can also set the number of clusters, and the method on which to base the clusters. The default number of terms used to describe the clusters is set at 5. That number may be too small to be able to label the clusters effectively, and it is recommended that this default be increased to 20 or more terms.

INVENTORY EXAMPLE

In this example, the charges from inventory ordered by different physicians in a hospital ER setting were examined for associations. There are hundreds of different charges listed. Text Miner was used to reduce the number of categories to ten with their identifications listed in Table 1 for the identified clusters.

Table 1. Text Miner Clusters from Charges

Cluster Number	*Descriptive Terms*	*Freq*	*Label*
1	+ low, + instruction, dc, th/pro/, intravenous, injectons, intravenous th/pro/, cbc, panel, insertion, lock, hep/saline, hep/saline lock insertion, p/visit, chemo, not, thpy, infus, chemo p/visit	1237 0	IV charges
2	+ monitor, cardiac, cardiac monitoring, pulse, oxymetry, daily, ox, ox monitoring, hour, holter, + transport, rn, simple	764	Heart monitoring
3	+ dress, complex, simple, triage, triage complex, nursing re-assessment, + nurse, re-assessment	3703	Bandagin g
4	ct, no, only, spine, wo, ct-head, dx, cervic-ap/lateral, ct-pelvic, ct-abd, lumbar-ap/lateral, thoracic-xray, regular, ct-chest, e.r., ct-spine, lamp, slit, abg, sacrum-xray, contrast, cervical-wo	1442	More complex X-ray and test
5	+ 4, iv, + 3, count, cell, glucose, fingerstick, + 1, + attempt, + tube, + 2, mini-neb, >, + 5, + 6, w/diff, hold, atrovent, protein/glucose, neb, alb/atrovent, csf, albuterol	1080	Diabetes and asthma
6	+ bone, xray-routine, + view, abd, xray, xray-portable, shoulder, + strap, acute, acute series-3, regular-ct, hip, w/pa, unilateral, rib, unilateral w/pa chest, xray-unilateral-two, xray-unilateral-tw	1308	X-ray series
7	oral, specimen, specimen collection, administration, med, collection, oxygen, exam, + point, pelvic exam assist, assist, admin, topical/rectal, topical/rectal med, + eye,	3655	Urinalysis and Pelvic

Cluster Number	Descriptive Terms	Freq	Label
	ear, irrigation, abd/vaginal		
8	urine, clean, clean catch, urinalysis, + wind, cleansing/irrigation, culture, pt, ptt, catch, catherization, cath, straight, foley, urine culture, stool, uretheral, magnesium, smear, wbcs, uric acid	1454	Bladder and kidney
9	extremity, troponin, test, qualitative, lipase, pregnancy, pregnancy test, screen, serum, amylase, rapid, + need, strep, rapid strep, prep, special, special needs, xray-ankle, xray-foot, + train	3894	Lab tests
10	im, injection, sq, ther/pro/d, antibiotic	1026	Antibiotic

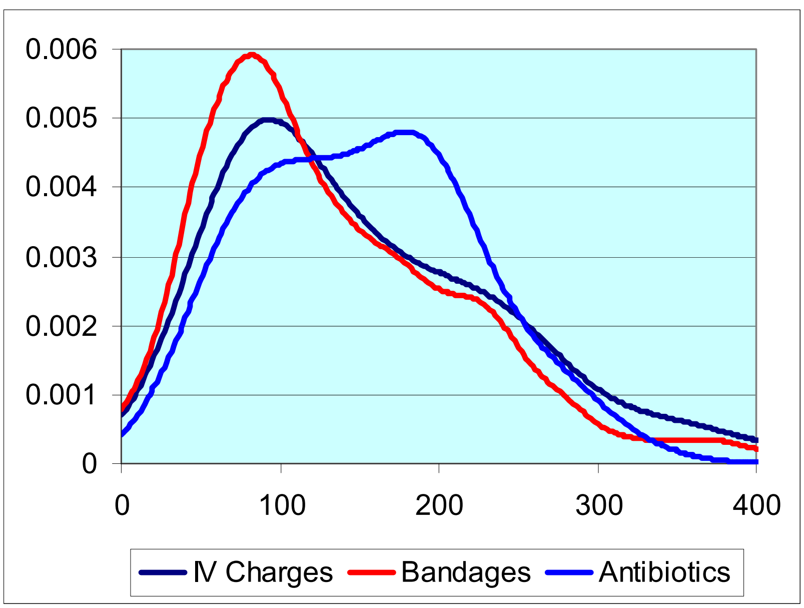
Using domain knowledge, a severity ranking was also provided (Table 2). The greater the severity, the higher the charge involved in treatment.

Table 2. Charge Categories in the ER

CHARGE CLUSTER	SEVERITY RANKING
IV charges	2
Heart monitoring	3
Bandaging	10
More complex X-ray and test	1
Diabetes and asthma	4
X-ray series	5
Urinalysis and Pelvic	9
Bladder and kidney	8
Lab tests	7
Antibiotic	6

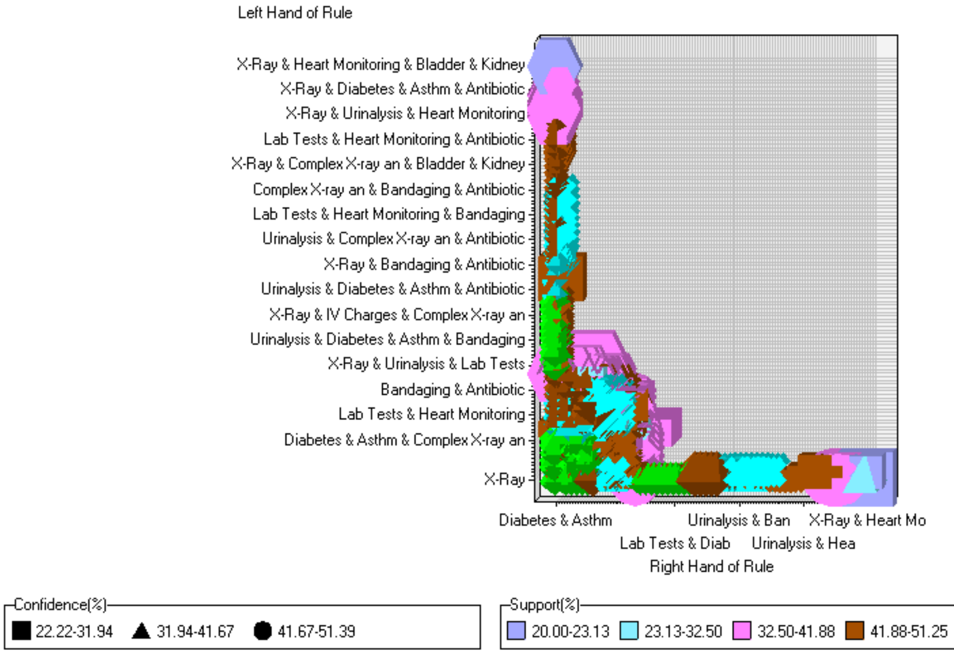
To investigate the level of severity, kernel density estimation (PROC KDE) was used to examine the patient length-of-stay in the ER (defined in minutes of time), and three clusters of charges (Figure 14). Note that patients receiving antibiotics will stay longer compared to patients receiving IV charges even though the severity ranking for IV charges is higher. It suggests (and requires additional analysis) that IV charges are given for more moderate patient complaints.

Figure 14. Length of ER Stay by Charge Cluster



The initial graph of associations is given in Figure 15.

Figure 15. Graph of Association Rules



There are already some clear indications of associations. All ten categories relate to diabetes and asthma. X-Ray clearly relates to all categories on the right-hand side. When the left-hand side is reduced to just one item, the graph becomes much simplified (Figure 16). This reduced graph demonstrates even more clearly that Diabetes and Asthma are associated with all categories of charges. Restricting the left-hand side to the category of antibiotic makes the results even clearer (Figure 17).

Antibiotics are generally associated with heart monitoring in some way. When the two are combined on the left-hand side, the number of associations reduces to 6 (Figure 18). Figure 19 gives the associations with Antibiotic and IV charges on the left-hand side. Again, they are all associated with heart monitoring. All of the associations have good confidence and support. Additional associations were examined relating to bandaging and X-rays (Figure 20).

Figure 16. Subset Graph

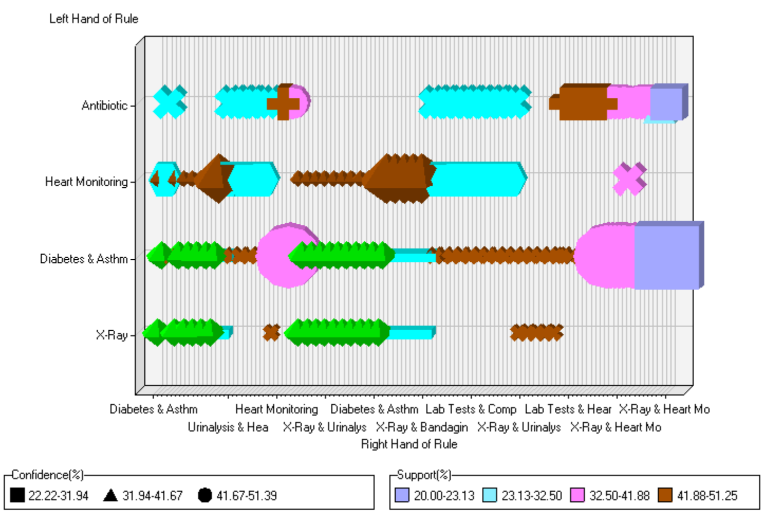


Figure 17. Associations With Antibiotics

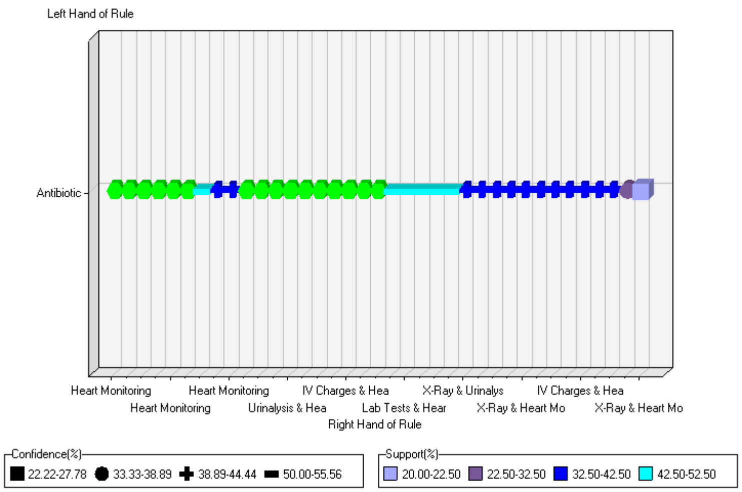


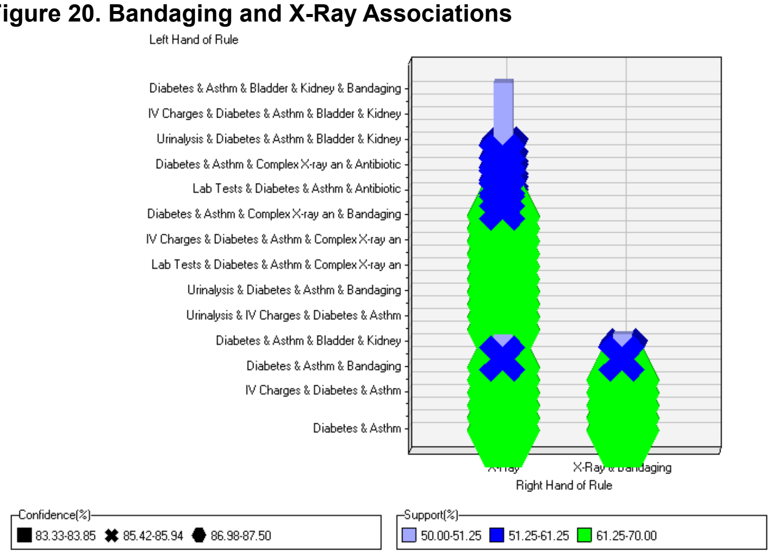
Figure 18. Associations with Antibiotics and Heart Monitoring

	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
1	3	1.04	50.00	83.33	5.00	Heart Monitoring & Antibiotic ==> Diabetes & Asthm
2	4	1.04	50.00	83.33	5.00	Heart Monitoring & Antibiotic ==> Urinalysis & Diabetes & Asthm
3	4	1.04	50.00	83.33	5.00	Heart Monitoring & Antibiotic ==> Lab Tests & Diabetes & Asthm
4	4	1.04	50.00	83.33	5.00	Heart Monitoring & Antibiotic ==> IV Charges & Diabetes & Asthm
5	4	1.04	50.00	83.33	5.00	Heart Monitoring & Antibiotic ==> Diabetes & Asthm & Complex X-ray an
6	4	1.04	50.00	83.33	5.00	Heart Monitoring & Antibiotic ==> Diabetes & Asthm & Bandaging

Figure 19. Associations with Antibiotics and IV Charges

	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
1	3	1.11	60.00	66.67	6.00	IV Charges & Antibiotic ==> Heart Monitoring
2	4	1.11	60.00	66.67	6.00	IV Charges & Antibiotic ==> Urinalysis & Heart Monitoring
3	4	1.11	60.00	66.67	6.00	IV Charges & Antibiotic ==> Lab Tests & Heart Monitoring
4	4	1.11	60.00	66.67	6.00	IV Charges & Antibiotic ==> Heart Monitoring & Complex X-ray an
5	4	1.11	60.00	66.67	6.00	IV Charges & Antibiotic ==> Heart Monitoring & Bandaging
6	4	1.11	50.00	55.56	5.00	IV Charges & Antibiotic ==> Heart Monitoring & Diabetes & Asthm
7	4	1.11	40.00	44.44	4.00	IV Charges & Antibiotic ==> X-Ray & Heart Monitoring
8	4	1.11	40.00	44.44	4.00	IV Charges & Antibiotic ==> Heart Monitoring & Bladder & Kidney

Figure 20. Bandaging and X-Ray Associations



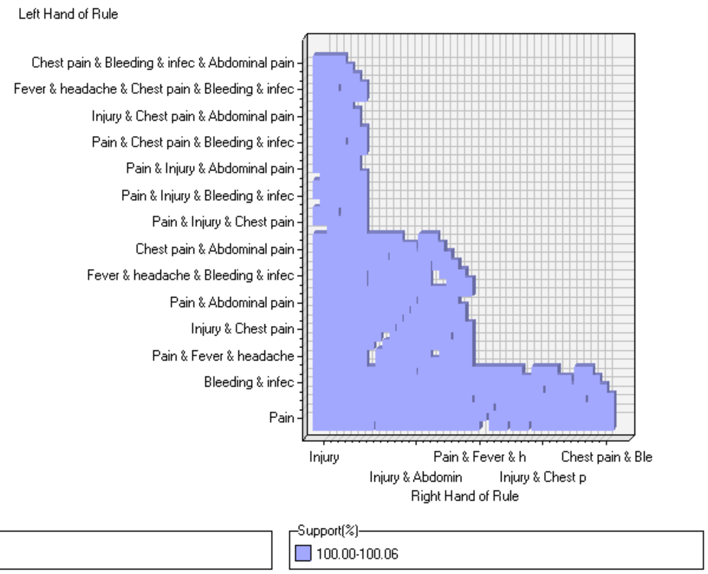
Note that X-ray is almost standard across many of the rules, but rarely with single categories on the left-hand side. The strongest association remains with diabetes and asthma. Similarly to a text clustering for charges, a text clustering was conducted for patient complaints. They tended to fall into a total of 6 categories (Table 3).

Table 3. Patient Complaints in the ER

Clusters
Fever and headache
Bleeding and infection
Injury
Pain
Abdominal pain
Chest pains

Again, with physicians as the ID variable and complaint clusters as the target, the association rules are given in Figure 21.

Figure 21. Association Rules for Patient Complaints



In this case, both life and support are at 100%. Effectively, the association computation has little value. The reason for this is simple. Each physician takes patients as they arrive in the ER, independent of the specific complaint. There is no choice. However, the charges are entirely at the discretion of the physician, providing choice.

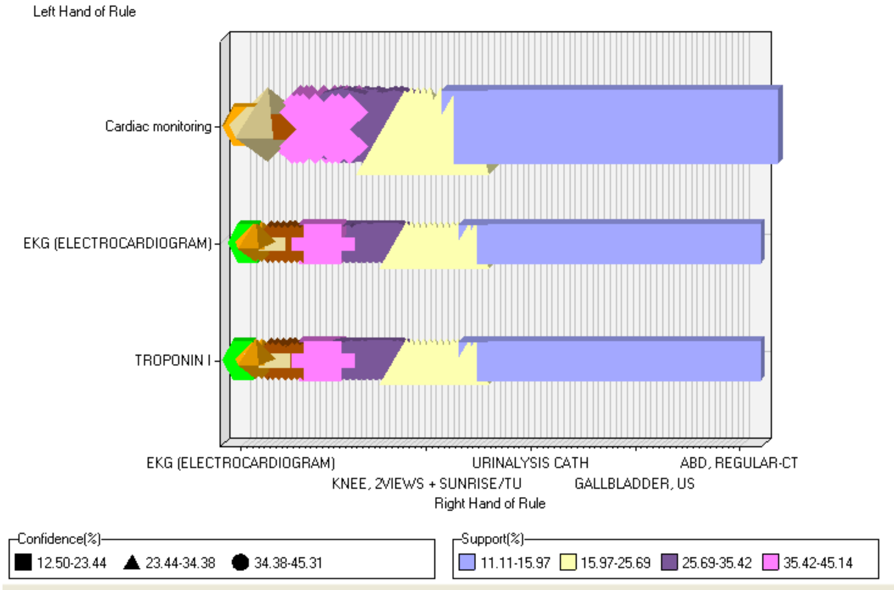
CONTRAST WITH ASSOCIATION RULES WITHOUT TEXT MINER

Using association rules on the inventory without first applying Text Miner resulted in difficulties. First, there were too many items in the list (approximately 30,000) with a Xeon dual-processor, 4-Gig RAM machine. To make the problem manageable, the data were sorted by MD and then the first 1000 items were sampled. The Association was still sluggish, resulting in no rules of relations>1. The sample was increased to 2500, again sorted by MD. This resulted in over 30,000 rules of level 2 and 3 (Figure 22).

	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
1	2	1.13	88.89	100.00	8.00	TROPONIN I ==> EKG (ELECTROCARDIOGRAM)
2	2	1.13	88.89	100.00	8.00	EKG (ELECTROCARDIOGRAM) ==> TROPONIN I
3	2	1.13	77.78	87.50	7.00	TROPONIN I ==> Cardiac monitoring
4	2	1.13	77.78	100.00	7.00	Cardiac monitoring ==> TROPONIN I
5	2	1.13	77.78	87.50	7.00	TROPONIN I ==> CBC W/ DIFFERENTIAL
6	2	1.13	77.78	100.00	7.00	CBC W/ DIFFERENTIAL ==> TROPONIN I
7	2	1.13	77.78	87.50	7.00	EKG (ELECTROCARDIOGRAM) ==> Cardiac monitoring
8	2	1.13	77.78	100.00	7.00	Cardiac monitoring ==> EKG (ELECTROCARDIOGRAM)
9	2	1.13	77.78	87.50	7.00	EKG (ELECTROCARDIOGRAM) ==> CBC W/ DIFFERENTIAL
10	2	1.13	77.78	100.00	7.00	CBC W/ DIFFERENTIAL ==> EKG (ELECTROCARDIOGRAM)
11	2	1.13	66.67	75.00	6.00	TROPONIN I ==> Pulse oxymetry
12	2	1.13	66.67	100.00	6.00	Pulse oxymetry ==> TROPONIN I
13	2	1.13	66.67	75.00	6.00	TROPONIN I ==> CT-HEAD, WO CONTRAST
14	2	1.13	66.67	100.00	6.00	CT-HEAD, WO CONTRAST ==> TROPONIN I
15	2	1.13	66.67	100.00	6.00	Pulse oxymetry ==> EKG (ELECTROCARDIOGRAM)
16	2	1.13	66.67	75.00	6.00	EKG (ELECTROCARDIOGRAM) ==> Pulse oxymetry
17	2	1.29	66.67	100.00	6.00	Pulse oxymetry ==> Cardiac monitoring
18	2	1.29	66.67	85.71	6.00	Cardiac monitoring ==> Pulse oxymetry
19	2	1.13	66.67	100.00	6.00	PREGNANCY TEST, QUALITATIVE ==> INJECTONS INTRAVENOUS TH/PRO
20	2	1.13	66.67	75.00	6.00	INJECTONS INTRAVENOUS TH/PRO/ ==> PREGNANCY TEST, QUALITATIV
21	2	1.13	66.67	75.00	6.00	EKG (ELECTROCARDIOGRAM) ==> CT-HEAD, WO CONTRAST
22	2	1.13	66.67	100.00	6.00	CT-HEAD, WO CONTRAST ==> EKG (ELECTROCARDIOGRAM)
23	2	1.10	66.67	85.71	6.00	Cardiac monitoring ==> CBC W/ DIFFERENTIAL
24	2	1.10	66.67	85.71	6.00	CBC W/ DIFFERENTIAL ==> Cardiac monitoring
25	2	1.13	66.67	100.00	6.00	CT-HEAD, WO CONTRAST ==> CHEST, XRAY-ROUTINE
26	2	1.13	66.67	75.00	6.00	CHEST, XRAY-ROUTINE ==> CT-HEAD, WO CONTRAST
27	2	1.13	55.56	100.00	5.00	Wound cleansing/irrigation ==> TROPONIN I
28	2	1.13	55.56	62.50	5.00	TROPONIN I ==> Wound cleansing/irrigation

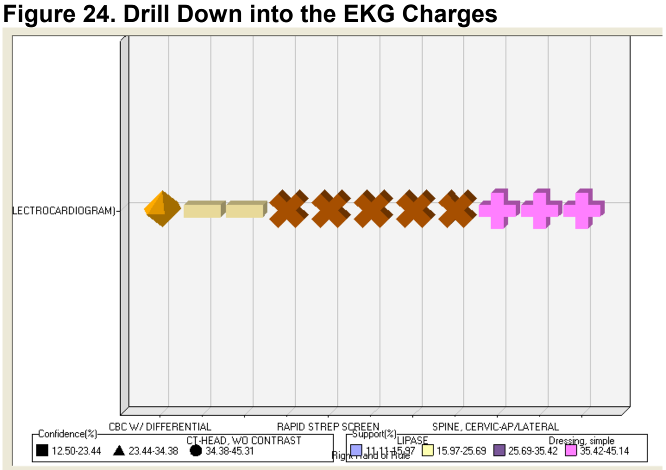
This is too many rules to depict graphically. Therefore, the rules will be reduced, starting with those most prominent having to do with suspected heart conditions (Troponin, cardiac monitoring, and EKG). The Results in the Association Node allow for the user to construct a subset table (figure 23).

Figure 23. Graph of Limited Rule Set



The result is still very dense. Some drill down into the graph is provided in Figure 24.

Figure 24. Drill Down into the EKG Charges



Note that the association with the highest level of confidence and support is with a CBC order (ie complete blood count). Limited to just these three charges on the left hand side, there are still 240 rules of level 2.

In the text mining, the CBC order falls into cluster 9, lab tests and the EKG falls in category 2, heart monitoring. Of the 407 patients with charges in the heart monitoring category, 324 (80%) also have charges in the lab cluster. In contrast, the confidence of the association of CBC to cardiac monitoring is equal to 85% but there are only 6 transactions containing both. Therefore, clustering can provide similar confidence levels but based on far more transactions. However, the association of cardiac monitoring with a blood culture is only 14%. Text Miner incorporates both into the two text clusters.

CONCLUSION

The combination of Text Miner with the Association Node in Enterprise Miner can provide meaningful intelligence, particularly with more limited computer resources. Association rules are most effective when the number of choices is small; however, in many situations the choices are in the thousands, or tens of thousands. Text Miner can combine different inventory items into more general categories that can then be used in the Association Node.

CONTACT INFORMATION

Patricia B. Cerrito
Department of Mathematics 502-852-6826
University of Louisville Fax: 502-852-7132
Louisville, KY 40292 Email:pcerrito@louisville.edu