

Link Analysis Using SAS® Enterprise Miner™

Ye Liu, Taiyeong Lee, Ruiwen Zhang, and Jared Dean

SAS Institute Inc.

ABSTRACT

The newly added Link Analysis node in SAS® Enterprise Miner™ visualizes a network of items or effects by detecting the linkages among items in transactional data or the linkages among levels of different variables in training data or raw data. This node also provides multiple centrality measures and cluster information among items so that you can better understand the linkage structure. In addition to the typical linkage analysis, the node also provides segmentation that is induced by the item clusters, and uses weighted confidence statistics to provide next-best-offer lists for customers. Examples that include real data sets show how to use the SAS Enterprise Miner Link Analysis node.

INTRODUCTION

Link analysis is a popular network analysis technique that is used to identify and visualize relationships (links) between different objects. The following questions could be nontrivial:

- Which websites link to which other ones?
- What linkage of items can be observed from consumers' market baskets?
- How is one movie related to another based on user ratings?
- How are different petal lengths, width, and color linked by different, but related, species of flowers?
- How are specific variable levels related to each other?

These relationships are all visible in data, and they all contain a wealth of information that most data mining techniques cannot take direct advantage of. In today's ever-more-connected world, understanding relationships and connections is critical. Link analysis is the data mining technique that addresses this need.

In SAS Enterprise Miner, the new Link Analysis node can take two kinds of input data: transactional data and non-transactional data (training data or raw data). The node can explore the relationships among transactional items and determine item clusters similar to how social network analysis determines communities. The node can also discover connections among levels of different variables in training data or raw data and produce data segmentations.

You can use the Link Analysis node for both transactional data and non-transactional data. (Non-transactional data are first converted to transactional data.) The basic steps of link analysis are as follows: First the Link Analysis node analyzes the (transformed) transactional data to define association rules or sequence rules. Then the node transforms the rules into a graph that consists of nodes and links. Next the node calculates a variety of centrality measures and detects item clusters. Finally the node scores the data: for transactional data, it produces either a next-best-offer list or customer segmentation information; for non-transactional data, it provides data segmentation information.

This paper first explains the concepts of association discovery, sequence discovery, multiple centrality measures and clustering coefficient measure, and item clusters. Then it shows how the Link Analysis node incorporates these concepts in analyzing transactional data. The paper also shows how you can adapt non-transactional data to the link analysis framework. Finally, examples illustrate how to use the Link Analysis node to analyze Netflix data and Fisher's Iris data.

LINK ANALYSIS CONCEPTS

The following subsections describe the main concepts in link analysis. For more information, see the *SAS Enterprise Miner: Reference Help*.

ASSOCIATION DISCOVERY

Association discovery is the identification of items that occur together in a particular event or record. This technique is also known as market basket analysis. Online transaction processing systems often provide the data sources for association discovery. Association discovery rules are based on frequency counts of the number of times that items occur alone and in combination in the database. The rules are expressed as "if item A is part of an event, then item B is also part of the event X% of the time." These rules are denoted as A->B.

The rules should not be interpreted as a direct causation, but as an association between two or more items. Association analysis does not create rules about repeating items, such as “if item A is part of an event, then another item A is also part of the event X% of the time.” In association analysis, it does not matter whether an individual customer buys one or multiple units of item A; only the presence of item A in the market basket is relevant. However, identifying creditable associations between two or more different items can help a business technologist make decisions such as when to distribute coupons, when to put a product on sale, or how to present items in store displays.

The following list shows some hypothetical association discovery rules:

- If a customer buys shoes, then 10% of the time he also buys socks.
- A grocery chain might find that 80% of all shoppers who purchase a bag of tortilla chips also buy a jar of salsa.
- When “do-it-yourselfers” buy latex paint, they also buy rollers 85% of the time.
- Of investors who hold an equity index fund, 40% also have a growth fund in their portfolio.

These example rules have a left-hand side (antecedent) and a right-hand side (consequent). In each rule in the preceding list, the first product is the antecedent item and the second product is the consequent item. Both sides of an association rule can contain more than one item.

The confidence factor, expected confidence, level of support, and lift statistics are four important evaluation criteria of association discovery:

- The strength of an association is defined by its *confidence factor*, which is the percentage of cases in which a consequent appears given that the antecedent has occurred.
- *Expected confidence* is equal to the number of consequent transactions divided by the total number of transactions.
- The *level of support* represents how frequently the combination occurs in the market basket (database).
- *Lift* is equal to the confidence factor divided by the expected confidence. Lift is a factor by which the likelihood of the consequent increases given an antecedent.

SEQUENCE DISCOVERY

Sequence discovery goes one step further than association discovery by taking into account the ordering of the relationships among items (the rules additionally imply a timing element). For example, rule $A \implies B$ implies that event B occurs after event A occurs.

Here are two hypothetical sequence rules:

- Of those customers who currently hold an equity index fund in their portfolio, 15% of them will open an international fund in the next year.
- Of those customers who purchase a new computer, 25% of them will purchase a laser printer in the next month.

CENTRALITY MEASURES AND THE CLUSTERING COEFFICIENT MEASURE

The centrality of a node in a network is a measure of its structural importance. There are many centrality measures in graph theory:

- **Degree Centrality:** Degree centrality represents a form of relative importance of a node to a network. It is simply the (in- or out-) degree of a node. The degree of a node v in an undirected graph is the number of links that connect to node v . The out-degree of a node in a directed graph is the number of links that start at this node and connect to another node; the in-degree is the number of links that start at another node and connect to this node. The terms degree, in-degree, and out-degree are interchangeable for an undirected graph.

- **Influence Centrality:** Influence centrality is a generalization of degree centrality that considers the link and node weights of adjacent nodes (C1) in addition to the link weights of nodes that are adjacent to the adjacent nodes (C2). The metric C1 is referred to as first-order influence centrality, and the metric C2 is referred to as second-order influence centrality. Let w_{uv} define the link weight for link (u, v), and let w_u define the node weight for node u. Let δ_u represent the list of nodes that are connected to node u (that is, its neighbors); this list is called the adjacency list. For directed graphs, the neighbors are the out-links. The general formula for influence centrality is:

$$C_1(u) = \frac{\sum_{v \in \delta_u} w_{uv}}{\sum_{v \in N} w_v}$$

$$C_2(u) = \sum_{v \in \delta_u} C_1(v)$$

- **Closeness Centrality:** Closeness centrality is the reciprocal of the average of the shortest paths (geodesic distances) to all other nodes. Closeness centrality can be thought of as a measure of how long it would take information to spread from a particular node to other nodes in the network.
- **Betweenness Centrality:** Betweenness centrality counts the number of times a particular node (or link) occurs on shortest paths between other nodes. Betweenness can be thought of as a measure of the control a node (or link) has over the communication flow among the rest of the network nodes. In this sense, the nodes (or links) that have high betweenness are the gatekeepers of information because of their relative location in the network.
- **Eigenvector Centrality:** Eigenvector centrality is an extension of degree centrality, in which centrality points are awarded for each neighbor. Because not all neighbors are equally important, a connection to an important node should contribute more to the centrality score than a connection to a less important node. This is the basic idea behind eigenvector centrality. Eigenvector centrality of a node is proportional to the sum of the centrality scores of all nodes that are connected to it.
- **Hub Centrality and Authority Centrality:** Hub centrality and authority centrality are used to rank the importance of web pages. Certain web pages (called hubs) are important because they point to many important pages (called authorities). A good hub node is one that points to many good authorities. A good authority node is one that is pointed to by many good hub nodes. Hub centrality and authority centrality apply only to directed graphs, which means that they are calculated only if input data are transactional data that have a sequence variable in link analysis.
- **Clustering Coefficient:** The clustering coefficient for a node is the number of links between the nodes within its neighborhood divided by the number of links that could possibly exist between them.

DETECTING ITEM CLUSTERS

Item-cluster detection partitions a graph into item clusters such that the links within the item clusters' subgraphs are more densely connected than the links between the item clusters.

LINK ANALYSIS APPLIED TO TRANSACTIONAL DATA

The Link Analysis node processes transactional data in the following steps:

1. It discovers association or sequence rules in transactional data and applies them.

Figure 1 shows how the confidence factor, level of support, expected confidence, and lift statistics are calculated.

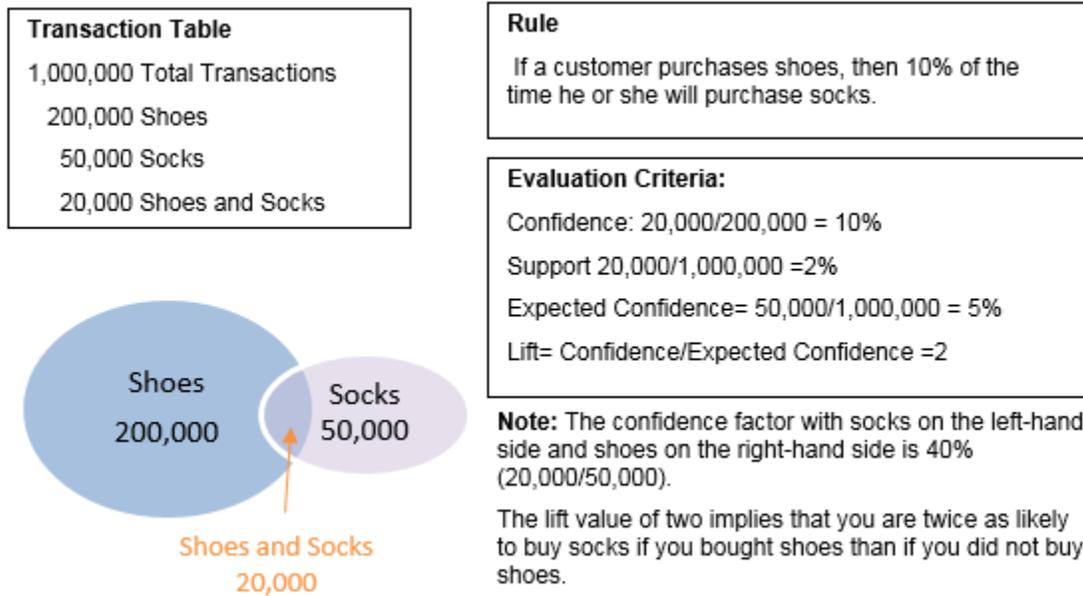


Figure 1. Association Discovery Statistics Example

A credible rule has a large confidence factor, a large level of support, and a lift greater than 1. Rules that have a high level of confidence but little support should be interpreted with caution.

- The node transforms the rules data into graph data (nodes and links).

From the association or sequence discovery in previous step, the transaction count for each item and for each two-item association or sequence rule are obtained. For the shoes and socks example, the two types of transaction count data are shown in Table 1 and Table 2.

Item	Count
Shoes	200,000
Socks	50,000
...	...

Table 1. One-Item Transaction Count Table

Two-Item Rules	Count
Shoes -> Socks	20,000
... ->

Table 2. Two-Item Transaction Count Table

The one-item transaction count data in Table 1 are transformed into node data straightforwardly as shown in Table 3.

Node	Node Weight
Shoes	200,000
Socks	50,000
...	...

Table 3. Node Data

The two-item association rules are transformed into **undirected** links data between two target items as shown in Table 4.

From	To	Link Weight
Shoes	Socks	20,000
...

Table 4. Links Data

Note: If a sequence variable exists, the two-item sequence rules are transformed into **directed** links data.

Figure 2 is a visualization of the graph data.



Figure 2. Link Graph

3. The node calculates multiple centrality measures and detects item clusters from the link graph.
4. It scores the transactional data.

Depending on the score property that is specified, the Link Analysis node produces either a next-best-offer list or customer segmentation information for scoring.

- Next-best-offer list for transactional data

The Link Analysis node uses weighted confidence to construct a next-best-offer list. Say you have some new customers' transaction data as shown in Table 5.

Customer	Items
1	A
1	B
2	A
2	C
3	...

Table 5. Example Customer Transaction Data

The link analysis produces an association or sequence rules table. Table 6 shows association rules; for sequence rules, '->' in the Rules column are replaced with "=>".

Rules	Count	Confidence
A->B	ct1	conf1
A->C	ct2	conf2
A->E	ct3	conf3
B->C	ct4	conf4
B->D	ct5	conf5
C->D

Table 6. Association Rules Table

Now you want to recommend items to the new customers based on the association or sequence rules.

Table 5 shows that customer 1 has A and B in his or her basket. If you look in Table 6 for rules whose left-hand item is A or B, you find a rules set that consists of A->B, A->C, A->E, B->C, and B->D. This rules set has items B, C, D, and E on the right-hand side; these are candidates for recommendation. Because item B

is already in the customer's basket, it is removed from the list of candidates. To determine which candidate the customer is mostly likely to want, you use a measure called weighted confidence for those candidates. The calculation formula is shown in Table 7.

Candidates	Weighted Confidence
C	$(ct2*conf2+ ct4*conf4)/(ct2+ct4)$
D	$ct5*conf5/ct5$
E	$ct3*conf3/ct3$

Table 7. Weighted Confidence Calculation Table

Finally, C, D, and E are recommended to customer 1, and they are ordered by weighted confidence. If a recommended item is already in the customer's basket, it is removed from the candidates.

- Customer segmentation information

The Link Analysis node provides customer segmentation information to help you determine different types of customers based on their transaction history.

Item intensity is used to segment customers. At the end of item clusters detection, a node could have links that connect to multiple item clusters. The intensity of a node that belongs to item cluster k is computed as the sum of the weights of links that connect item cluster k divided by the total link weights of the node.

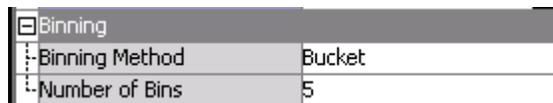
For each item j in a customer's basket, the intensity of j (a node) that belongs to an item cluster k is I_{jk} . The segment score for this customer is calculated as follows:

$$S = \arg \max_k \sum_j I_{jk}$$

LINK ANALYSIS APPLIED TO NON-TRANSACTIONAL DATA

Now that you are familiar with transactional data, you can learn how the Link Analysis node handles data that have a role of raw or train.

1. The node transforms raw or training data to transactional data.
 - a. In the Link Analysis node's Property panel, you specify binning properties, as shown in Display 1, in order to bin the interval variables (including the target variable).



Binning	
Binning Method	Bucket
Number of Bins	5

Display 1. Binning Properties in Property Panel of Link Analysis Node

- b. Each level of binned interval variables and categorical variables is treated as a unique item. Then each observation (ID) has a basket of unique items from different input variables. The raw or training data have been transformed to transactional data.

Table 8 shows some simple raw data.

Obs	X1	X2	X3
1	2.5	70	A
2	4.6	40	C
3	3.6	56	B
4	4.3	45	C

Table 8. Simple Raw Data

If you specify bucket binning and three bins, you get binned levels as shown in Table 9.

Obs	X1	X2	X3
1	BIN_X1_1	BIN_X2_3	A
2	BIN_X1_3	BIN_X2_1	C
3	BIN_X1_2	BIN_X2_2	B
4	BIN_X1_3	BIN_X2_1	C

Table 9. Transform Raw Data to Binned Data

If you treat each observation as a transaction basket, Table 8 can be transformed into transactional data as shown in Table 10.

Customer	Items
1	BIN_X1_1
1	BIN_X2_3
1	A
2	BIN_X1_3
2	BIN_X2_1
2	C
3	BIN_X1_2
3	BIN_X2_2
3	B
4	BIN_X1_3
4	BIN_X2_1
4	C

Table 10. Transformed Data

2. The node follows step 2 in the section “Link Analysis Applied to Transactional Data.”

Now you have a linkage graph of items, which are the interval variables’ binned levels and the categorical variables’ levels. You can apply all the analyses to the graph just as you did in the transactional data session.

3. It scores the non-transactional data.

For raw or training data, the Link Analysis node provides observation segmentation scores. Item intensity is used to segment observations. You can see each observation as a customer and each input variable value as an item in the customer’s basket. Then you apply the same segment score formula for the transactional data to raw or training data and get the segmentation information you need.

EXAMPLES

TRANSACTIONAL DATA AS INPUT

The data in this example are modified from the training data in Bennett and Lanning (2007). Netflix provided the training data set, which consists of 100,480,507 ratings that 480,189 users gave to 17,770 movies. Each rating is a quadruplet of the form $\langle user, movie, date-of-grade, grade \rangle$. The *user* and *movie* fields are integer IDs, and *grades* are from 1 to 5 (integral) stars.

The data is subsetted by applying the criterion $grade \geq 3$ and then dropping the *date-of-grade* variable.

First, use the following steps to configure the Link Analysis node:

1. Set the variable roles.

Right-click on the Ids (input data set) node and select **Edit Variables**. Assign the roles as shown in Display 2.

Name 	Role	Level
grade	Frequency	Interval
movie	Target	Nominal
user	ID	Nominal

Display 2. Variable Roles

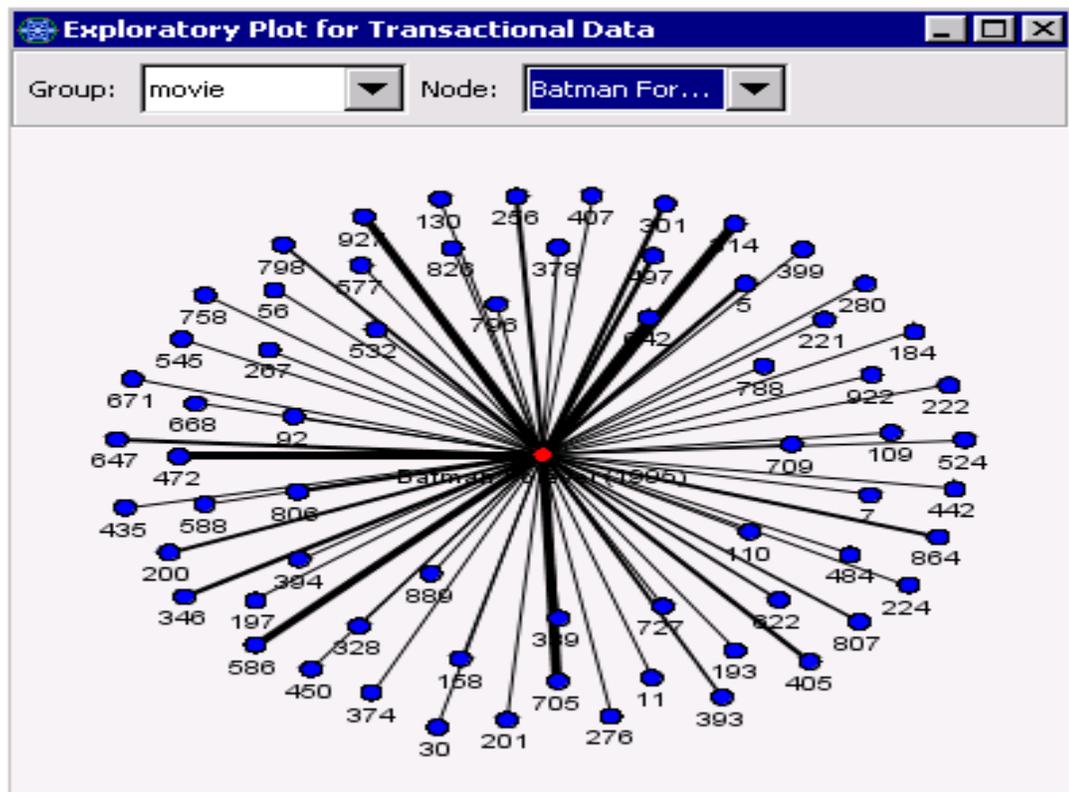
- In the Train section of the Link Analysis node's Property panel, modify the **Association Support Percentage** from 5 to 0.5, as shown in Display 3. The change ensures that at least two association rules can be generated. This property sets the minimum support counts to be 0.5% of the largest item frequency for one-item sets and two-item sets.

Train	
Variables	...
Minimum Confidence(%)	5
Association Support Type	Percent
Association Support Percent	0.5
Association Support Count	5
Sequence Property Settings	...

Display 3. Train Properties Window

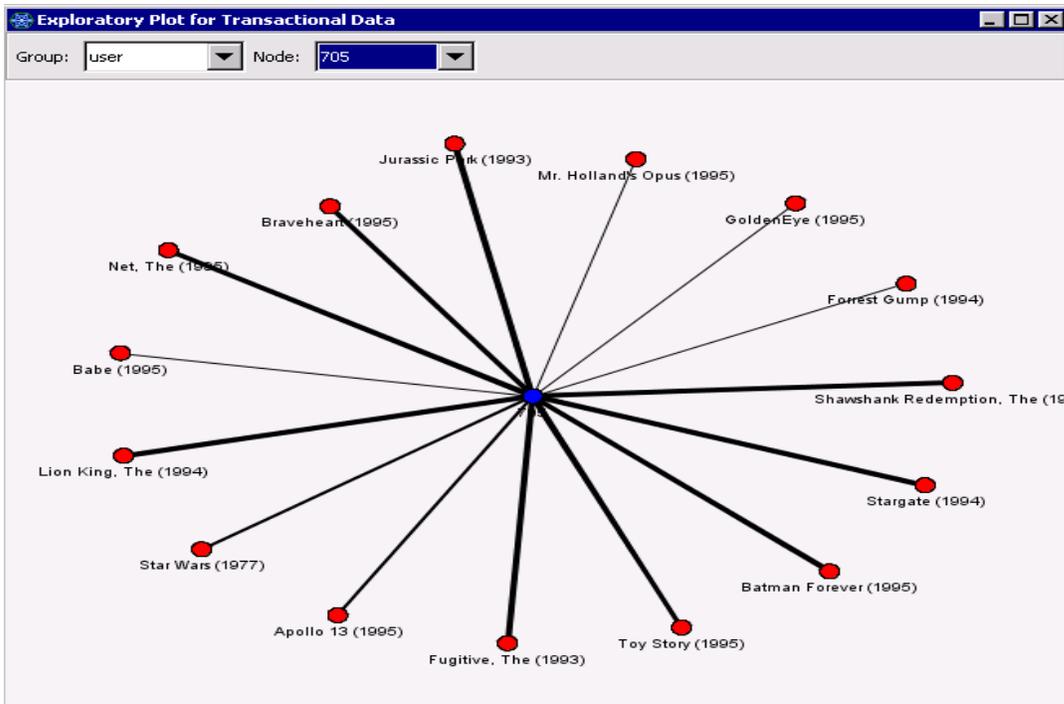
- Ensure that the following statement is included in your Project Start Code editor window:


```
options mprint;
```
- Keeping all other properties as default, run the Link Analysis node.
- After the run completes successfully, open the Results window.
- Select **View->Plots** and click on the exploratory plot for transactional data to get some general idea of the data. When you select **movie** from the **Group** list and **Batman Forever (1995)** from the **Node** list, you can see how different users grade the movie Batman Forever (1995) as shown in Display 4. The thickness of a link represents a user's rating for the movie.



Display 4. How Users Grade for the Movie Batman Forever (1995)

You can see that user 705 gives a very high rating for Batman Forever in Display 4. If you want to see what other movies user 705 likes, you can select user in the **Group** list and 705 in the **Node** list. Display 5 shows the result.



Display 5. How User 705 Grades All the Movies He or She Watched

Now you can examine the link analysis processing of transactional data by following the steps that are described in the section "LINK ANALYSIS APPLIED TO TRANSACTIONAL DATA." The following steps show how you can examine those steps in this example:

1. Discover the associations.

Open the log and look for the PROC ASSOC run. Output 1 shows that the minimum support level for one item sets is 181 ($=36364 \times 0.5\%$) and that the minimum support level for two item sets is 11 ($=2261 \times 0.5\%$).

```

----- Potential 1 item sets = 1506 -----
Counting items, records read: 100000
Counting items, records read: 153834
Number of customers: 65884
Support level for item sets: 181
Maximum count for a set: 36364
Sets meeting support level: 121
Megs of memory used: 1.02

----- Potential 2 item sets = 7260 -----
Counting items, records read: 100000
Counting items, records read: 153834
Maximum count for a set: 2261
Sets meeting support level: 395
Megs of memory used: 1.54

```

Output 1. Partial Log Contents

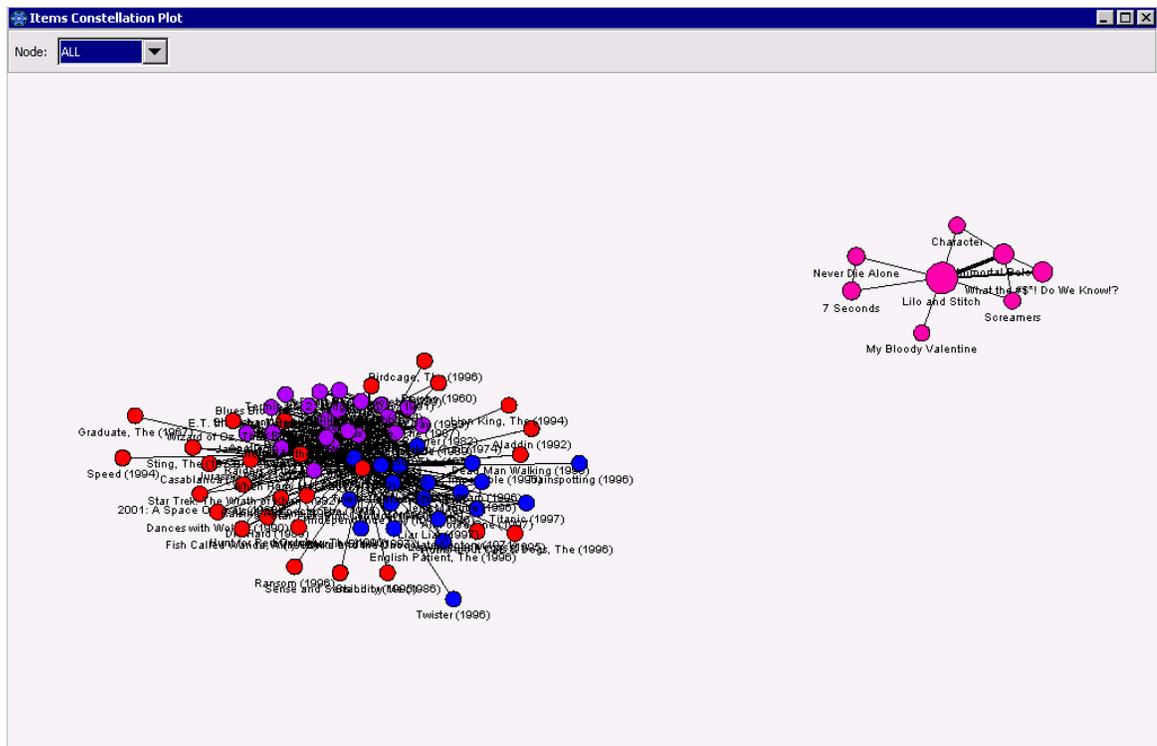
The association discovery yields the expected confidence, confidence, support, and lift statistics as shown in Display 6.

Table: Rule Statistics by Rule ID						
Rule ID	Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count	Rule
1	0.85	94.54	0.68	111.62	450.00	Return of the Jedi (1983) ==> Star Wars (1977)
2	0.85	93.43	0.50	110.31	327.00	Empire Strikes Back, The (1980) ==> Star Wars (1977)
3	0.85	92.95	0.32	109.75	211.00	Star Trek: The Wrath of Khan (1982) ==> Star Wars (1977)
4	0.85	92.08	0.28	108.72	186.00	Aladdin (1992) ==> Star Wars (1977)
5	0.85	92.02	0.30	108.65	196.00	Hunt for Red October, The (1990) ==> Star Wars (1977)
6	0.85	91.77	0.32	108.36	212.00	Jurassic Park (1993) ==> Star Wars (1977)
7	0.85	91.71	0.29	108.28	188.00	Lion King, The (1994) ==> Star Wars (1977)
8	0.61	91.55	0.30	149.67	195.00	Hunt for Red October, The (1990) ==> Raiders of the Lost Ark (1981)
9	0.61	91.54	0.28	149.66	184.00	Speed (1994) ==> Raiders of the Lost Ark (1981)
10	0.85	91.04	0.39	107.49	254.00	Terminator, The (1984) ==> Star Wars (1977)
11	0.61	90.91	0.32	148.62	210.00	Jurassic Park (1993) ==> Raiders of the Lost Ark (1981)
12	0.61	90.71	0.31	148.29	205.00	Die Hard (1988) ==> Raiders of the Lost Ark (1981)
13	0.85	90.42	0.36	106.76	236.00	Blade Runner (1982) ==> Star Wars (1977)
14	0.85	90.27	0.31	106.58	204.00	Die Hard (1988) ==> Star Wars (1977)
15	0.61	89.96	0.38	147.08	251.00	Terminator, The (1984) ==> Raiders of the Lost Ark (1981)

Display 6. Association Statistics

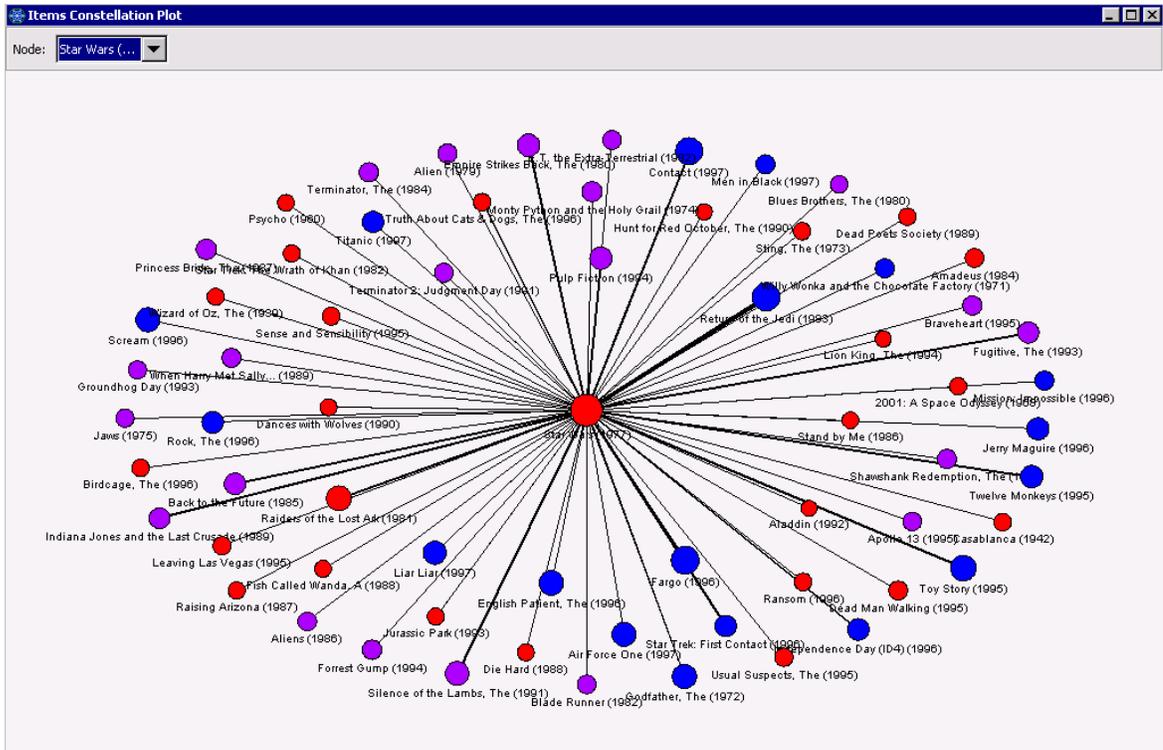
2. Transform the rules data into graph data.

The rules data yield nodes and links data in an items constellation plot as shown in Display 7. Color represents which item cluster a node belongs to.



Display 7. Items Constellation Plot

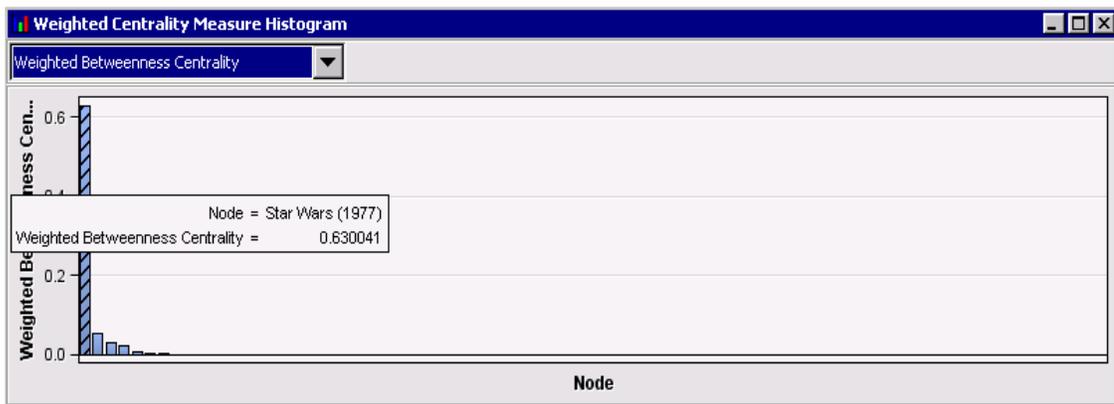
If you are interested in particular node (for example, Star Wars (1977) in this case), you can select the node from the Node list. Then all nodes connected to the node you specified are displayed, as shown in Display 8.



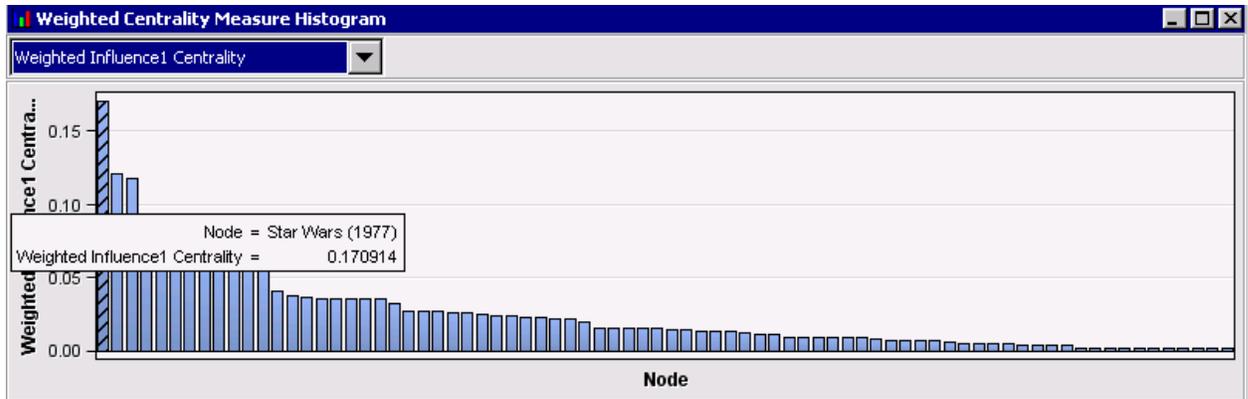
Display 8. Nodes That Star Wars (1977) Connects To

3. View the multiple centrality measures and item clusters.

The node data for the items constellation plot contain all the centrality and item-cluster statistics that are computed from the graph. You can also check the weighted centrality measure histogram in the Results window.



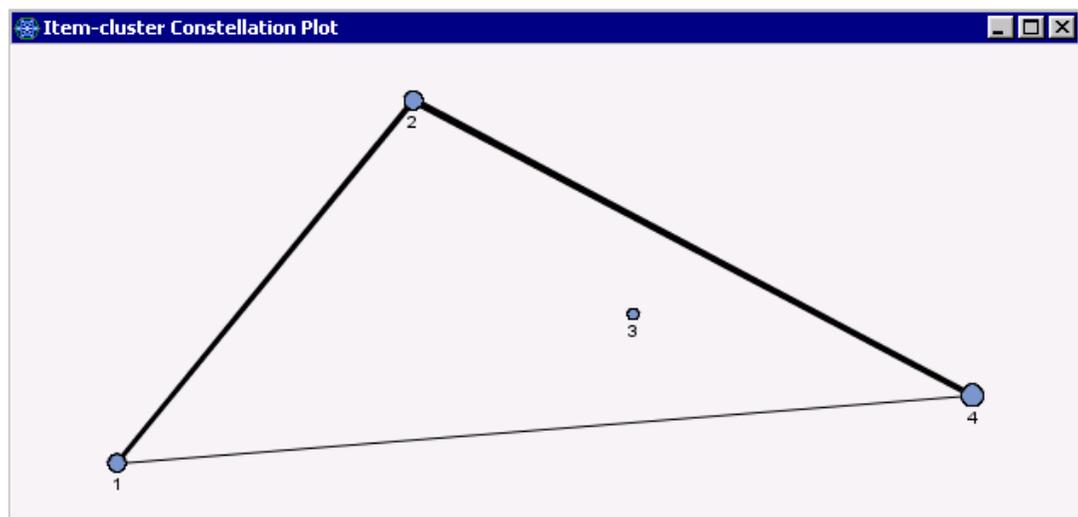
Display 9. Weighted Betweenness Centrality Measure Histogram



Display 10. Weighted Betweenness Centrality Measure Histogram

You can see that the movie Star Wars (1977) has the highest weighted betweenness centrality, which means that the load and importance of the node is the highest of all the nodes. It also has the highest weighted first-order influence centrality, which means that it is the most influential node of all the other nodes.

The item-cluster constellation plot in Display 11 shows that four item clusters have been detected and that the item clusters 2 and 4 have the strongest inter-item-cluster bonds. Item cluster 3 has no inter-item-cluster bond with any other item clusters.



Display 11. Item-cluster Constellation Plot

4. Score your data.

When you score your data, you can generate either a next-best-offer list or segmentation information.

- Next-best-offer list

If you set **Generate Recommendation Table** to **Yes** (the default setting) in the Score property window, the next-best-offer list is displayed in the score output. You can export the full recommendation table, or the filtered recommendation table that results from specifying properties such as "Top N," "Minimum Confidence (%)," and "Criterion Relation," as shown in Display 12.

Recommendation	
Generate Recommendation Table	Yes
Top N	3
Minimum Confidence(%)	.
Criterion Relation	AND
Export Recommendation Table	Filtered Recommendation Table

Display 12. Recommendation Table Properties Window

If you specify the settings shown in Display 12, a filtered recommendation table is displayed, as shown in Display 13. The table contains the top three next-best offers for each customer ID.

ID Variable	Next Best Offer	Confidence	Rank
1	E.T. the Extra-Terrestrial (1982)	56.74261	1
1	Jerry Maguire (1996)	51.21059	2
1	Mission: Impossible (1996)	49.1109	3
2	Return of the Jedi (1983)	67.56392	1
2	Raiders of the Lost Ark (1981)	56.99736	2
2	Silence of the Lambs, The (1991)	54.96026	3
3	Star Wars (1977)	94.53782	1
3	Raiders of the Lost Ark (1981)	66.80672	2
3	Toy Story (1995)	64.4958	3
4	Empire Strikes Back, The (1980)	67.70773	1
4	Raiders of the Lost Ark (1981)	66.35461	2
4	Return of the Jedi (1983)	65.74312	3
5	Pulp Fiction (1994)	64.25532	1
5	Terminator, The (1984)	59.76817	2
5	Terminator 2: Judgment Day (1991)	57.62803	3

Display 13. Filtered Recommendation Table

- Segmentation Information

If you set **Generate Recommendation Table** to **No** in the Score property window (see Display 14), segmentation information is displayed in the Score output's training data, as shown in Display 15.

Score	
Recommendation	
Generate Recommendation Table	No
Top N	1
Minimum Confidence(%)	.
Criterion Relation	AND
Export Recommendation Table	Full Recommendation Table

Display 14. Score Property Window Setting for Segmentation Information Generation

	user	Segment ID
1	1.0	4.0
2	2.0	1.0
3	3.0	2.0
4	4.0	1.0
5	5.0	4.0
6	6.0	2.0
7	7.0	4.0
8	8.0	2.0
9	9.0	1.0
10	10.0	4.0
11	11.0	4.0

Display 15. Score Output of Segmentation Information

NON-TRANSACTIONAL DATA AS INPUT

As a non-transactional data example, you can use DMAIRIS data in the SAMPSIO library. The data set contains or Fisher-Anderson's Iris data. You can generate this data set through the SAS Enterprise Miner **Help** menu.

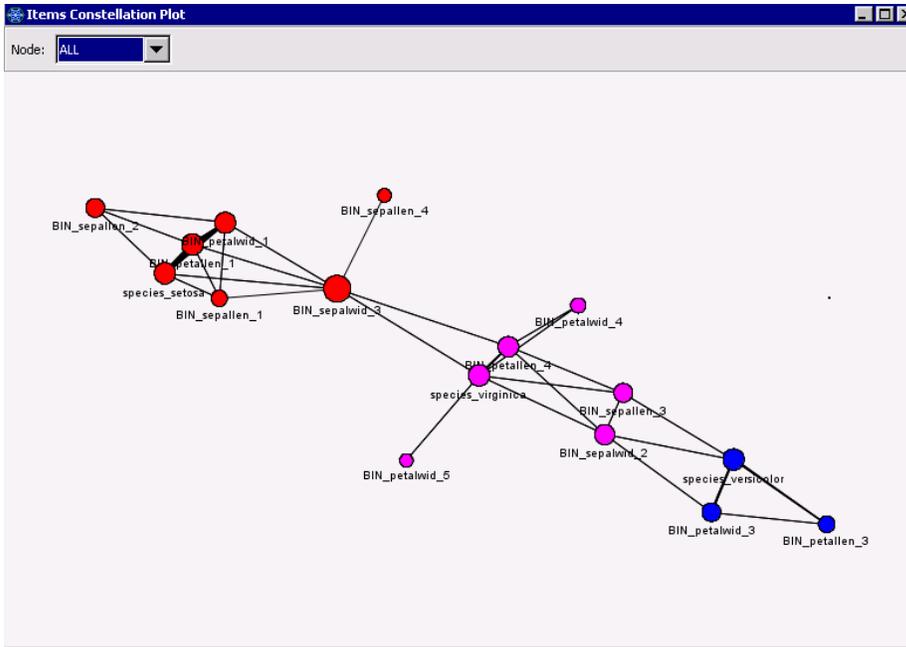
The data set consists of 50 samples from each of three species of iris (*Iris setosa*, *Iris virginica*, and *Iris versicolor*). Four features were measured from each sample: the length and the width of the sepals and of the petals, in centimeters.

In this example, a 35% minimum confidence and 30% association support are specified in the Train property window, and default values are kept for other properties, as shown in Display 16. You might want to tune the training properties and rerun the node after you check the items constellation plot in the results.

Train	
Variables	...
Minimum Confidence(%)	35
Association Support Type	Percent
Association Support Percentage	30.0
Association Support Count	5
Sequence Property Settings	...
<input checked="" type="checkbox"/> Binning	
Binning Method	Bucket
Number of Bins	5

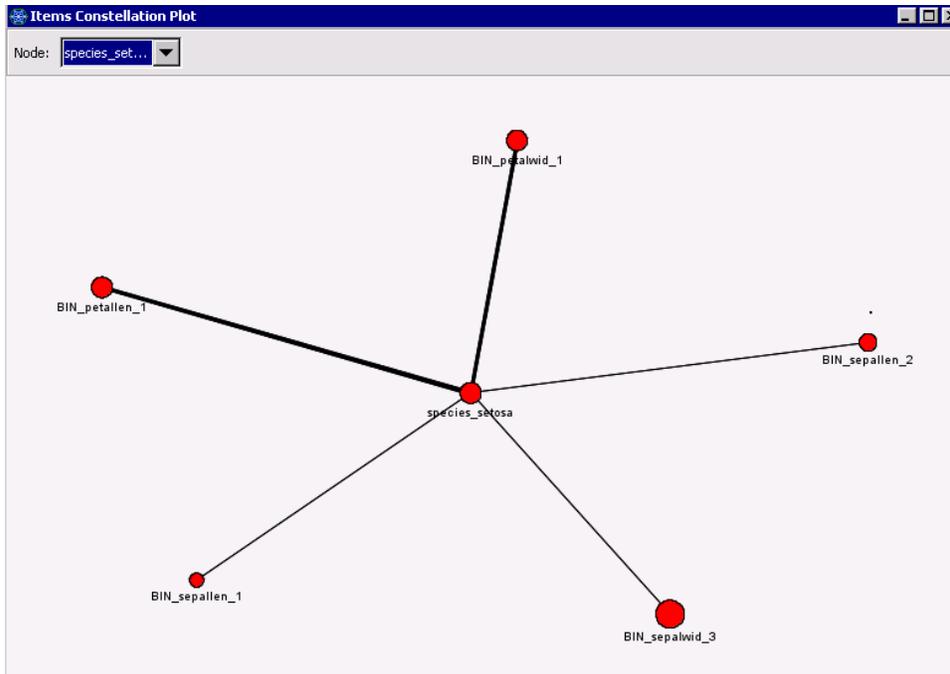
Display 16. Train Property Window Setting

Display 17 shows the items constellation plot. You can see that three item clusters are detected and are represented in three different colors.



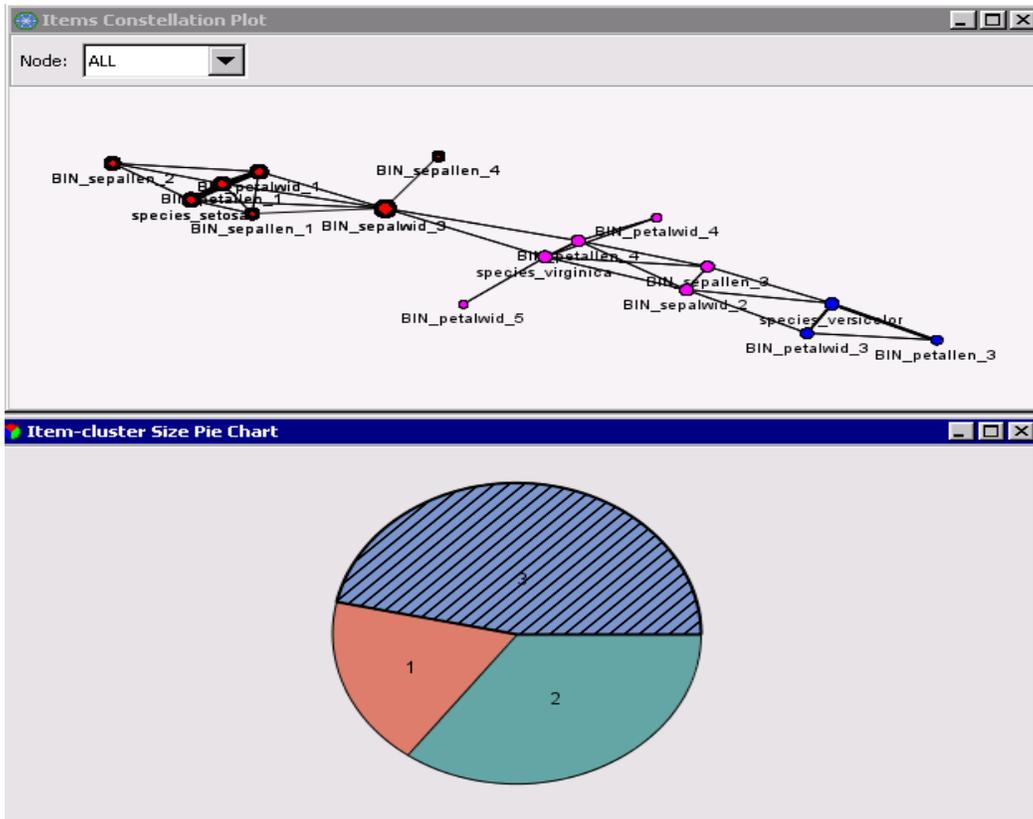
Display 17. Items Constellation Plot

If you select `species_setosa` in the **Node** list, the window displays the plot shown in Display 18. The plot indicates that the setosa species is associated with features such as petal length less than 21.8 (BIN_petalen_1), petal width less than 5.8 (BIN_petalwid_1), sepal length less than 57.4 (BIN_sepalen_1 and BIN_sepalen_2), and sepal width from 29.6 to 34.4 (29.6 not included, BIN_sepalwid_3). This information is useful when you do the segmentation for the flowers. For each binned variable level's range, you can refer to the score code.



Display 18. Features Species Setosa Is Associated With

In the Results window of Link Analysis node, some visualizations interact. For example, when you click item cluster 3 in the Item-Cluster Size Pie Chart, the nodes in item cluster 3 are highlighted in the Items Constellation Plot. See Display 19.



Display 19. Interaction between Item-Cluster Size Pie Chart and Items Constellation Plot

For raw data, no score properties are applied. The segmentation information is added to the score output data as shown in Display 20. The Segment ID column includes the scoring results of segmentation.

	species	Sepal Length in mm.	Sepal Width in mm.	Petal Length in mm.	Petal Width in mm.	Segment ID
1	setosa	51.0	35.0	14.0	2.0	3.0
2	setosa	46.0	31.0	15.0	2.0	3.0
3	setosa	50.0	36.0	14.0	2.0	3.0
4	setosa	54.0	39.0	17.0	4.0	3.0
5	setosa	46.0	34.0	14.0	3.0	3.0
6	setosa	50.0	34.0	15.0	2.0	3.0
7	setosa	49.0	31.0	15.0	1.0	3.0
8	setosa	54.0	37.0	15.0	2.0	3.0
9	setosa	48.0	34.0	16.0	2.0	3.0
10	setosa	48.0	30.0	14.0	1.0	3.0
11	setosa	58.0	40.0	12.0	2.0	3.0
12	setosa	54.0	39.0	13.0	4.0	3.0

Display 20. Score Output Data

SUMMARY

This paper shows how to use Link Analysis node in SAS Enterprise Miner to analyze transactional and non-transactional data. Netflix Prize data is modified as an example of transactional data. The results from the Link Analysis node show how movies are linked to each other by users' ratings and show some other interesting results. Another example uses non-transactional data, the famous Fisher's Iris data set. From the results, you can see how

different petal lengths, width, and color are linked by different but related species of Iris flowers. In addition to the core link analysis, a unique segmentation method and a next-best-offer tool have been shown based on the link analysis result.

REFERENCES

Bennett, J. and Lanning, S. 2007. The Netflix Prize. Proceedings of KDD Cup and Workshop 2007, San Jose, CA. Aug 12, 2007.

ACKNOWLEDGMENTS

The authors would like to thank Anne Baxter for editing this paper.

CONTACT INFORMATION

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

Ye Liu
ye.liu@sas.com

Taiyeong Lee
taiyeong.lee@sas.com

Ruiwen Zhang
ruiwen.zhang@sas.com

Jared Dean
jared.dean@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.