

# Mining an Online Auctions Data Warehouse

David Ulmer

Under the guidance of Assistant Professor Sung-Hyuk Cha  
Pace University

## Abstract

*This report evaluates a marketing opportunity for an auction business in the online auction industry. The proposal is specific to the sale of art, antiques and collectibles and represents sales conducted as English auctions (as opposed to other forms of auction or retail). The analysis will consist of a Market Basket Analysis (MBA) for evaluating cross-promotional sales and marketing opportunities in the business. The taxonomy of the product hierarchy contains collecting categories and the analysis will determine if customers purchase across collecting categories. The input to the MBA is a sales dataset that has been denormalized to contain repeating groups of collecting category sales data. The MBA utilizes the Apriori algorithm, which is a association rule discovery procedure. It is a tabulation algorithm that determines the frequent itemsets and generates the association rules by finding relationships between the items in the dataset. The report concludes with the association rules determined as part of the analysis.*

## 1. Introduction

Many businesses attempt to identify customer purchasing patterns to understand if these patterns offer additional business opportunities. In the auction industry, the customers tend to be classified as collectors (i.e., a person who makes a collection) who have a loyalty to a particular product category (e.g., coin or stamp collectors). The main goal and concept description of this report is:

***To mine the database of buying activity to determine if collectors have an affinity to buy across product categories.***

The answer to this question could enable important marketing decisions in advertising, targeted email or cross-promoted online sales. This proposal utilizes a Market Basket Analysis [5] to determine if customers purchase across collecting categories. This same analysis is well known by the example of customers who purchase diapers also buy beer on Thursdays. MBA is an association rule mining technique that is used to find items that are most frequently purchased together. It uses the Apriori algorithm [1] and generates the appropriate association rules by pruning the itemset space by filtering on the (3) metrics of support, confidence and improvement (support and confidence are sometimes called coverage and accuracy).

Agrawal et al. introduced the Apriori algorithm for the purpose of efficient association rule mining [1,3]. The concept of association rules was introduced in 1993 [2]. The Apriori algorithm is an efficient association discovery algorithm that filters item sets by incorporating item constraints (support). Agrawal et al. actually introduced several algorithms in addition to Apriori (e.g., AprioriTid, AprioriHybrid) that were oriented to mining association rules for large databases. They also demonstrated performance improvements over the then accepted AIS and SETM algorithms. Although efficient, recent progress has been made with other algorithms that use the power of parallel processors to split the problem into multiple parts and other algorithmic improvements. An example of one of these improved algorithms is Dynamic Itemset Counting [4].

## 2. Input and Output

There are more attributes in the sales dataset than actually required for this analysis. This is to enable the additional statistical analysis beyond the Market Basket Analysis (see research extensions). The data file is in a relational data format and the metadata for the attributes is as follows:

<b>Normalized Attribute</b>	<b>Description</b>	<b>Format</b>
BuyerID	Unique identifier for the individual buyer	Numeric – integer
Date	Date of sale	MM/DD/YYYY
Category	Product category of item	Text – nominal
Price	Hammer price in USD	Numeric - integer

Since the results of the analysis may be used for targeted marketing campaigns, the actual buyer is required so grouped data would not suffice. The analysis will be conducted over the past year (2001) since the earlier data is not thought to be representative of the problem space. The date attribute will be used to filter the data, but also could be used to do seasonal pattern analysis. The category attribute is nominal data representing the category code. The category code is also a codified value. The price attribute could be nominal for this analysis, but it is represented as an integer value to enable further statistical analysis (beyond the scope of this report). The source of the data is a corporate sales data warehouse.

To simplify and prepare the data for analysis, the relational data needs to be denormalized since the purchases could occur over various dates (multi-category collectors may not make purchases on the same date because of scarcity of items to buy). The denormalized file will contain the BuyerID followed by a repeating group of (13) categories (i.e., all of the categories the buyer has ever purchased within). The denormalized file structure is:

<b>Denormalized Attribute</b>	<b>Description</b>
BuyerID	Unique identifier for the individual buyer
Category1	Product category of 1st item purchased
....	....
Category13	Product category of the 13th item purchased

In this case, the data was transformed to its denormalized state through a series of queries and operations in Microsoft Access and Excel. As an alternative to the denormalized data structure, the data could also be coded in the ARFF format in order to utilize the Apriori algorithm/code that is available from the Weka web site.

### *Knowledge Representation*

Several different types of output are appropriate for a Market Basket Analysis. For example, a tabular representation is a two-dimensional co-occurrence matrix. The rows and columns would represent the frequent item set combinations (i.e., categories purchased together most often) and the data would be the frequency count of the combination. The co-occurrence matrix can be used to calculate additional two-dimensional tables representing the metrics of support, confidence and improvement.

The co-occurrence matrix can be “mined” by running the Apriori algorithm across the itemset space, which will determine if any strong relationships exist between the items in the dataset. This process can transform the matrix into association rules for easier interpretation.

## **3. Algorithms**

### *Evaluation Methods*

Association rule mining is a procedure that looks for relationships between items in a dataset. The relationships can be determined through statistical methods such as correlation analysis or tabulation procedures such as Market Basket Analysis. MBA is useful for determining which items tend to be purchased together. In this case it will be used to identify multi-category collectors. MBA is typically expressed as association rules in the form “If *Condition* then *Result*” (e.g., “If Item Y is purchased then Item Z is purchased”).

The MBA process consists of:

**A. Choosing the right items (using a taxonomy or other filtering methods)**

**B. Running a cross-tabulation to create the co-occurrence matrix**

**C. Run Apriori and determine the quality of the association rules by calculating the support, confidence and improvement**

The Apriori algorithm is an iterative process, which successively processes each k-sized itemset (i.e., an itemset that contains k items) until no larger itemsets can be generated. The specific steps of the algorithm are:

1. Set  $K=1$
2. Calculate all k-sized itemsets
3. Calculate the support for all of the candidate itemsets – filter the itemsets based on a minimum support metric
4. Join all k-sized itemsets to generate candidate itemsets for size  $K+1$

5. Set  $K = K + 1$
6. Repeat steps 3 through 5 until no larger itemsets can be formed
7. Generate the final set of itemsets by creating the union of all  $k$ -sized itemsets

The Apriori algorithm utilizes the three metrics to find “strong” rules by pruning the itemset space. The good or *useful* rules (rules with a quality course of action) are the ones we want to find, and ignore the *trivial* (obviously known rules) and *inexplicable* rules (rules that seem to have no explanation and no course of action) that Apriori will also find.

Support is the percentage of records containing the item combination compared to the total number of records [7]. Confidence is the ratio of the number of transactions with all the items in the rule to the number of transactions with just the items in the condition. An example of confidence is:

*If A & B => C has an 80% confidence means, if when a collector buys A and B, in 80% of the cases, the collector also bought C.*

Improvement measures how much better a rule is at predicting a result than just assuming the result. When it is  $> 1$ , the rule is better at predicting the result than random chance.

#### D. Generate association rules

- a) Finding the most frequent combinations of item sets
- b) Define condition and result (for the conditional association rules)

\* \* \* \*

The Market Basket Analysis was done using a data mining software product from Megaputer called PolyAnalyst [6]. PolyAnalyst (version 4.5) is a powerful tool for discovering patterns and knowledge hidden in your data. This proposal used the MBA features of PolyAnalyst to determine the association rules for this dataset. Microsoft Excel was also used to calculate the co-occurrence, support and confidence matrices. The results were:

#### A. Choosing the right items (using a taxonomy or other filtering methods)

2001 sales data across (13) collecting categories

#### B. Running a cross-tabulation to create the co-occurrence matrix

Category	CAT01	CAT02	CAT03	CAT04	CAT05	CAT06	CAT07	CAT08	CAT09	CAT10	CAT11	CAT12	CAT13
CAT01	474	75	55	55	40	71	37	110	35	50	31	19	16
CAT02	75	814	68	77	108	185	93	155	39	81	79	35	13
CAT03	55	68	1169	170	95	162	78	212	82	137	82	39	14
CAT04	55	77	170	2049	113	185	124	198	99	118	104	54	13
CAT05	40	108	95	113	993	317	99	183	34	96	153	50	12
CAT06	71	185	162	185	317	1780	161	425	76	173	212	82	13
CAT07	37	93	78	124	99	161	1073	149	49	97	113	89	12
CAT08	110	155	212	198	183	425	149	2543	142	392	139	92	19
CAT09	35	39	82	99	34	76	49	142	783	99	32	25	4

<b>CAT10</b>	50	81	137	118	96	173	97	392	99	1647	68	50	8
<b>CAT11</b>	31	79	82	104	153	212	113	139	32	68	610	63	8
<b>CAT12</b>	19	35	39	54	50	82	89	92	25	50	63	511	5
<b>CAT13</b>	16	13	14	13	12	13	12	19	4	8	8	5	121

### C. Run Apriori and determine the quality of the association rules by calculating the support, confidence and improvement

PolyAnalyst allows you to input and control each of the three parameters (support, confidence and improvement). Setting the minimum support allows you to prune the product groups (in our case categories) that Apriori finds. For instance, if the minimum support is set high, then only category groups found in large numbers of transactions together will be returned (which is more likely to be groups of 2 categories). Setting it lower makes it more likely that category groups with 3 or more categories will be returned. Confidence and improvement are used by PolyAnalyst to prune the resultant association rules. For this analysis, category groups of 2 categories are more important since they represent clear opportunities to cross-promote marketing for those collecting categories (typically, product groups of 3 or more are used for organizing a store or shelf).

In addition to calculating support, confidence and improvement for each category group, PolyAnalyst also calculates the p-value (probability). The lower the value, the less likely the group appeared by random chance. The co-occurrence count is also displayed on the same line displayed as the label called support.

The power of PolyAnalyst is the ability to quickly run a different analysis with different support, confidence and improvement. If you run an analysis and it returns to many or to few groups, you can reset the minimum support and rerun the analysis. As a result, changing the minimum support will result in different groups being returned, which are all valid, but valid for different purposes. For example, as mentioned, a lower minimum support will tend to return groups with more products within them. This is particularly useful for product positioning.

The following matrices were calculated in Microsoft Excel to crosscheck the results of PolyAnalyst (in step D.)

#### Support

Category	CAT01	CAT02	CAT03	CAT04	CAT05	CAT06	CAT07	CAT08	CAT09	CAT10	CAT11	CAT12	CAT13
<b>CAT01</b>	4.5%	0.7%	0.5%	0.5%	0.4%	0.7%	0.4%	1.0%	0.3%	0.5%	0.3%	0.2%	0.2%
<b>CAT02</b>	0.7%	7.7%	0.6%	0.7%	1.0%	1.8%	0.9%	1.5%	0.4%	0.8%	0.8%	0.3%	0.1%
<b>CAT03</b>	0.5%	0.6%	11.1%	1.6%	0.9%	1.5%	0.7%	2.0%	0.8%	1.3%	0.8%	0.4%	0.1%
<b>CAT04</b>	0.5%	0.7%	1.6%	19.5%	1.1%	1.8%	1.2%	1.9%	0.9%	1.1%	1.0%	0.5%	0.1%
<b>CAT05</b>	0.4%	1.0%	0.9%	1.1%	9.4%	3.0%	0.9%	1.7%	0.3%	0.9%	1.5%	0.5%	0.1%
<b>CAT06</b>	0.7%	1.8%	1.5%	1.8%	3.0%	16.9%	1.5%	4.0%	0.7%	1.6%	2.0%	0.8%	0.1%
<b>CAT07</b>	0.4%	0.9%	0.7%	1.2%	0.9%	1.5%	10.2%	1.4%	0.5%	0.9%	1.1%	0.8%	0.1%
<b>CAT08</b>	1.0%	1.5%	2.0%	1.9%	1.7%	4.0%	1.4%	24.1%	1.3%	3.7%	1.3%	0.9%	0.2%
<b>CAT09</b>	0.3%	0.4%	0.8%	0.9%	0.3%	0.7%	0.5%	1.3%	7.4%	0.9%	0.3%	0.2%	0.0%
<b>CAT10</b>	0.5%	0.8%	1.3%	1.1%	0.9%	1.6%	0.9%	3.7%	0.9%	15.6%	0.6%	0.5%	0.1%
<b>CAT11</b>	0.3%	0.8%	0.8%	1.0%	1.5%	2.0%	1.1%	1.3%	0.3%	0.6%	5.8%	0.6%	0.1%

CAT12	0.2%	0.3%	0.4%	0.5%	0.5%	0.8%	0.8%	0.9%	0.2%	0.5%	0.6%	4.9%	0.0%
CAT13	0.2%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.2%	0.0%	0.1%	0.1%	0.0%	1.1%

**Confidence**

Category	CAT01	CAT02	CAT03	CAT04	CAT05	CAT06	CAT07	CAT08	CAT09	CAT10	CAT11	CAT12	CAT13
CAT01	100.0%	15.8%	11.6%	11.6%	8.4%	15.0%	7.8%	23.2%	7.4%	10.5%	6.5%	4.0%	3.4%
CAT02	9.2%	100.0%	8.4%	9.5%	13.3%	22.7%	11.4%	19.0%	4.8%	10.0%	9.7%	4.3%	1.6%
CAT03	4.7%	5.8%	100.0%	14.5%	8.1%	13.9%	6.7%	18.1%	7.0%	11.7%	7.0%	3.3%	1.2%
CAT04	2.7%	3.8%	8.3%	100.0%	5.5%	9.0%	6.1%	9.7%	4.8%	5.8%	5.1%	2.6%	0.6%
CAT05	4.0%	10.9%	9.6%	11.4%	100.0%	31.9%	10.0%	18.4%	3.4%	9.7%	15.4%	5.0%	1.2%
CAT06	4.0%	10.4%	9.1%	10.4%	17.8%	100.0%	9.0%	23.9%	4.3%	9.7%	11.9%	4.6%	0.7%
CAT07	3.4%	8.7%	7.3%	11.6%	9.2%	15.0%	100.0%	13.9%	4.6%	9.0%	10.5%	8.3%	1.1%
CAT08	4.3%	6.1%	8.3%	7.8%	7.2%	16.7%	5.9%	100.0%	5.6%	15.4%	5.5%	3.6%	0.7%
CAT09	4.5%	5.0%	10.5%	12.6%	4.3%	9.7%	6.3%	18.1%	100.0%	12.6%	4.1%	3.2%	0.5%
CAT10	3.0%	4.9%	8.3%	7.2%	5.8%	10.5%	5.9%	23.8%	6.0%	100.0%	4.1%	3.0%	0.5%
CAT11	5.1%	13.0%	13.4%	17.0%	25.1%	34.8%	18.5%	22.8%	5.2%	11.1%	100.0%	10.3%	1.3%
CAT12	3.7%	6.8%	7.6%	10.6%	9.8%	16.0%	17.4%	18.0%	4.9%	9.8%	12.3%	100.0%	1.0%
CAT13	13.2%	10.7%	11.6%	10.7%	9.9%	10.7%	9.9%	15.7%	3.3%	6.6%	6.6%	4.1%	100.0%

**D. Generate association rules**

*Output*

Three different analyses were run in PolyAnalyst to determine which produced the most appropriate association rules for this particular business opportunity. By evaluating the calculated support matrix, it is obvious that in general the category relationships have a low support. The three runs set support at .5, 1 and 2. No other runs produced any additional association rules. The results were:

**Run 1 -- Exploration parameters:**

Argument	Value
	Numeric values has been binarized
Minimum support, %	0.5
Minimum improvement	1
Minimum confidence, %	1

---

3 group(s) of associated products were found

group #1 contains 3 products  
p\_value: 0.000000 (log = -146.338357); support 100  
CAT11  
CAT05  
CAT06

---

group #2 contains 2 products  
 p\_value: 0.000000 (log = -17.876041); support 75  
 CAT02  
 CAT01

group #3 contains 2 products  
 p\_value: 0.000002 (log = -13.172577); support 89  
 CAT12  
 CAT07

**PRODUCT ASSOCIATION RULES**

	Filter:	Support	Confidence	Improvement
		> 0.50	> 1.00	> 1.00
CAT05, CAT06	-> CAT11	0.95%	31.55%	5.447
CAT11, CAT06	-> CAT05	0.95%	47.17%	5.003
CAT11, CAT05	-> CAT06	0.95%	65.36%	3.868
CAT01	-> CAT02	0.71%	15.82%	2.047
CAT02	-> CAT01	0.71%	9.21%	2.047
CAT07	-> CAT12	0.84%	8.29%	1.710
CAT12	-> CAT07	0.84%	17.42%	1.710

**Run 2 -- Exploration parameters:**

Argument	Value
	Numeric values has been binarized
Minimum support, %	1
Minimum improvement	1
Minimum confidence, %	1

2 group(s) of associated products were found

---

group #1 contains 2 products  
 p\_value: 0.000000 (log = -57.624735); support 153  
 CAT11  
 CAT05

group #2 contains 2 products  
 p\_value: 0.000061 (log = -9.707988); support 185  
 CAT06  
 CAT02

PRODUCT ASSOCIATION RULES

	Filter:	Support	Confidence	Improvement
		> 1.00	> 1.00	> 1.00
CAT05	-> CAT11	1.45%	15.41%	2.661
CAT11	-> CAT05	1.45%	25.08%	2.661
CAT02	-> CAT06	1.76%	22.73%	1.345
CAT06	-> CAT02	1.76%	10.39%	1.345

**Run 3** -- Exploration parameters:

Argument	Value
	Numeric values has been binarized
Minimum support, %	2
Minimum improvement	1
Minimum confidence, %	1

1 group(s) of associated products were found.

group #1 contains 2 products  
 p\_value: 0.000000 (log = -56.576025); support 317  
 CAT06  
 CAT05

PRODUCT ASSOCIATION RULES

	Filter:	Support	Confidence	Improvement
		> 2.00	> 1.00	> 1.00
CAT05	-> CAT06	3.01%	31.92%	1.889
CAT06	-> CAT05	3.01%	17.81%	1.889

*Interpretation of Association Rules*

The (3) runs of the MBA essentially produced six category groups (along with their related directional groups). One group contains 3 categories, which for this analysis is less interesting (because it is hard to cross-promote on the fact of a customer purchasing within 2 categories).

The relationships aren't particularly strong by evidence of the low support and moderately low confidence results. This indicates that there are minimal cross-promotion opportunities within this business. These results do support the anecdotal business opinion that customers of these types of items have a "collectors" mentality, which enforces a strong loyalty to a particular collecting category. To the extent that there are opportunities, the association rules can be interpreted as:

- If a client collects (purchases) CAT01, then they also collect CAT02. The confidence is 15.82% (meaning the chance of purchasing the CAT02 is 15.88%). Also, it is 2.047 (improvement) times more likely than random chance that this association rule is valid (i.e., statistically sound). The additional rule is the same (lower confidence at 9.21%), but opposite in the direction of the relationship.
- If a client collects CAT07, then they also collect CAT12. The rule of opposite direction is valid as well.
- If a client collects CAT05, then they also collect CAT11. The rule of opposite direction is valid as well.
- If a client collects CAT02, then they also collect CAT06. The rule of opposite direction is valid as well.
- If a client collects CAT05, then they also collect CAT06. The rule of opposite direction is valid as well.

#### 4. Research Extensions

Beyond the Market Basket Analysis, the dataset has additional attributes, which could be used for additional analysis. For example, since the dataset has a time dimension, seasonal pattern analysis could be done with additional statistical techniques. Also, given that the price is numeric, additional descriptive statistics could be calculated including an analysis of variance (ANOVA). Correlation coefficients could also be calculated between the combinations to determine statistical correlation between the purchasing patterns.

#### References

- [1] Rakesh Agrawal, and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules in large databases. *In Proceedings of the 20<sup>th</sup> International Conference on Very Large Data Bases*, vol. 2, pp. 478-499, Santiago, Chile, September 1994.
- [2] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami, Mining association rules between sets of items in large databases. *In Proceedings of the ACM SIGMOD International Conference on Management of Data*, Vol. 22, pp. 207-216, June 1993.
- [3] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and Inkeri Verkamo. Fast discovery of association rules. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, Chapter 12, AAAI Press, 1996.
- [4] Sergey Brin, Rajeev Motwani, Jeffrey Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. *In Proceedings of the ACM SIGMOD International Conference on Management of Data*, New York, May 1997. <http://www-db.stanford.edu/~sergey/dic.html> , [2002, April]

[5] Grant Bugher, Market Basket Analysis of the Sales Data for a Client of Cambridge Technology Partners, [http://www.megaputer.com/company/cases/cambridge\\_mba.php3](http://www.megaputer.com/company/cases/cambridge_mba.php3), [2003, March]

[6] Megaputer Intelligence, *Data Mining with PolyAnalyst*, <http://www.megaputer.com>, [2002, March]

[7] Ian Witten, and Eibe Frank. 2000. *Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations*, San Francisco, CA: Morgan Kaufmann.