MARKET BASKET ANALYSIS: UNDERSTANDING INDIAN CONSUMER BUYING BEHAVIOR OF SPAIN MARKET

Prof. Kavitha Venkatachari (kavitav@ibsindia.org)
IBS Business School, Powai, Mumbai, India

Abstract

Market Basket Analysis is a useful tool for retailers who want to better understand the relationships between the products that people buy. There are many tools that can be applied when carrying out Market basket analysis and the trickiest aspects to the analysis are setting the confidence and support thresholds in the apriori algorithm and identifying which rules are worth pursuing. Ultimately the key to MBA is to extract value from the transaction data by building up an understanding of the needs of the consumers. This type of information is invaluable if you are interested in marketing activities such as cross-selling or targeted campaigns. Retailers nowadays are having large amounts of data but poor in information extracted from that data. Big data is seen as a valuable resource and although the concept of data mining is still new and developing, companies in a variety of industries are relying on it for making strategic decisions. Such information can be used as a basis for decisions about marketing activity such as promotional support, inventory control and cross-sale campaigns. The main objective of the research paper is to see how different products in a mall interrelate and how to exploit these relations by marketing activities. Mining association rules from transactional data will provide us with valuable information about co-occurrences and co-purchases of products. Such information can be used as a basis for decisions about marketing activity such as promotional support, inventory control and cross-sale campaigns. The data is collected from the Indian consumers who purchase the products from Spain retail shops.

Business use of datamining

Data mining is commonly seen as a single step of a whole process called Knowledge Discovery in Databases (KDD). According to Fayyad et.al, 'KDD is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.' (Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, 1996) Data mining is a technique that encompasses a huge variety of statistical and computational techniques such as: association-rule mining, neural network analysis, clustering, classification, summarizing data and of course the traditional regression analyses.

Data mining gained popularity especially in the last two decades when advances in computing power provided us with the possibility to mine voluminous data. Extracting knowledge and hidden information from data using a whole set of techniques found its applications in various contexts. Knowledge discovery is widely used in marketing to identify and analyze customer groups and predict future behaviour. Data mining is an effective way to provide better service to customers and adjust offers according to their needs and motivations.

Companies nowadays are rich in vast amounts of data but poor in information extracted from that data. Big data is seen as a valuable resource and although the concept of data mining is still new and developing, companies in a variety of industries are relying on it for making strategic decisions. Facts that otherwise may go unnoticed can be now revealed by the techniques that sift through stored information. When applying mining tools and techniques we seek to find useful relationships, patterns and anomalies that can help managers make better business decisions.

Data mining tools perform analyses that are very valuable for business strategies, scientific research and getting to know your customers better. Managerial insights are no longer the only factor trusted when it comes to decision-making. Data driven decisions can lead to better firm performance.

Objective of the study

In the recent years analyzing shopping baskets has become quite appealing to retailers. Advanced technology made it possible for them to gather information on their customers and what they buy. The introduction of

electronic point-in sale increased the use and application of transactional data in market basket analysis. In retail business analyzing such information is highly useful for understanding buying behavior. Mining purchasing patterns allows retailers to adjust promotions, store settings and serve customers better.

Identifying buying rules is crucial for every successful business. Transactional data is used for mining useful information on co-purchases and adjusting promotion and advertising accordingly. The well-known set of rubber and pencil is just an example of an association rule found by data scientists.

The main objective of the thesis is to see how different products in a retail shop assortment interrelate and how to exploit these relations by marketing activities. Mining association rules from transactional data will provide us with valuable information about co-occurrences and co-purchases of products. Some shoppers may purchase a single product during a shopping trip, out of curiosity or boredom, while others buy more than one product for efficiency reasons.

Motive of the Research

It is very important for retailers to get to know what their customers are buying. Some products have higher affinity to be sold together and hence the retailer can benefit from this affinity if special offers and promotions are developed for these products. It is also important to the retailer to cut off products from the assortment which are not generating profits. Deleting loss-making, declining and weak brands may help companies boost their profits and redistribute costs towards aspects of the more profitable brands. (Kumar, 2009) This is yet another reason why data mining is seen as a powerful tool for many businesses to regularly check if they are selling too many brands, identify weak ones and possibly merge them with healthy brands. Data mining techniques are highly valued for the useful information they provide so that the retailer can serve customers better and generate higher profits.

1. Find products with affinity to be sold together.
2. Improve in-store settings and optimize product placement.
3. Improve layout of the catalogue of e-commerce site.
4. Control inventory based on product demand.

Literature Review

Data mining has taken an important part of marketing literature for the last several decades. Market basket analysis is one of the oldest areas in the field of data mining and is the best example for mining association rules.

Various algorithms for Association Rule Mining (ARM) and Clustering have been developed by researchers to help users achieve their objectives. Rakesh Agrawal and Usama Fayyad are one of the pioneers in data mining. They account for a number of developed algorithms and procedures.

According to Shapiro, rule generating procedures can be divided into procedures that find quantitative rules and procedures that find qualitative rules. (Rakesh Agrawal, Ramakrishnan Srikant) elaborate on the concept of mining quantitative rules in large relational tables. Quantitative rules are defined in terms of the type of attributes contained in these relational tables. Attributes can be either quantitative (age, income, etc.) or categorical (certain type of a product, make of a car). Boolean attributes are such attributes that can take on one of two options (True or False, 1 or 0). They are considered a special case of categorical attributes. The authors call this mining problem the Quantitative Association Rules problem. An example of a generated quantitative rule is :

If ((Age : [30…39]) + (Married : Yes)) → (Number of cars = 2)

The example combines variables that have quantitative and boolean attributes.

(S. Prakash, R.M.S. Parvathi, 2011) propose a qualitative approach for mining quantitative association rules. The nature of the proposed approach is qualitative because the method converts numerical attributes to binary attributes.

However, finding qualitative rules is of main interest in this analysis. These rules are most commonly represented as decision trees, patterns or dependency tables. (Gregory Piatetsky-Shapiro, William Frawley, 1991) The type of attributes used for mining qualitative rules is categorical.

(Rakesh Agrawal, Tomasz Imielinski, Arun Swami, 1993) is one of the first published papers on association rules that proposes a rule mining algorithm that

discoversqualitative rules with no restriction for Boolean attributes. The authors test the effectiveness of the algorithm by applying it to data obtained from a large retailing company.

Association rules found application in many research areas such as: market basket analysis, recommendation systems, intrusion detection etc.

In marketing literature market basket analysis has been classified into two models: explanatory and exploratory. First, exploratory models will be thoroughly explained in this paper as they are of higher relevance for the research and after that an explanation of explanatory models will be given. The main idea behind exploratory models is the discovering of purchase patterns from POS (point-of-sale) data. Exploratory approaches do not include information on consumer demographics or marketing mix variables. (KatrinDippold, Harald Hruschka, 2010) Methods like association rules (Rakesh Agrawal, Sirkant Ramakrishnan, 1994) or collaborative filtering (Andreas Mild, Thomas Reutterer, 2003) summarise a vast amount of data into a fewer meaningful rules or measures. Such methods are quite useful for discovering unknown relationships between the items in the data. Moreover, these methods are computationally simple and can be used for undirected data mining. However, exploratory approaches are not appropriate for forecasting and finding the cause-roots of complex problems. They are just used to uncover distinguished cross-category interdependencies based on some frequency patterns for items or product categories purchased together.

A typical application of these exploratory approaches is identifying product category relationships by simple association measures. Pairwise associations are used to compare entities in pairs and judge which entity is prefered or has greater amount of some quantitative property. (Julander, 1992) compares the percentage of shoppers buying a certain product and the percentage of all total sales generated by this product. By making such comparisons, one can easily find out the leading products and what is their share of sales. Examining which the leading products are for consumers is extremely important since a large number of shoppers come into contact with these specific product types every day.

As the departments with leading products generate much in-store traffic, it is crucial to use this information for placing other specific products nearby. The paper by Julander also shows how combinatory analysis can be used to study the patterns of cross-buying between certain brands or product groups: for instance, what is the percentage of shoppers that buy products A+C, but not B or what is the percentage of shoppers that buy only A. It also deals with the probabilities that shoppers will purchase from one, two or more departments in a single visit in the store.

Another significant stream of research in the field of exploratory analysis is the process of generating association rules. Substantial amount of algorithms for mining patterns from market basket data have been proposed. From the co-operative work of Rakesh Agrawal and Ramakrishnan Srikant they present two new algorithms for discovering large item sets in databases, namely Apriori and AprioriTid. These two algorithms are similar with regard to the function that is used to determine the candidate itemsets, but the difference is that the AprioriTID does not use the database for counting support after the first pass (first iteration) while Apriori makes multiple passes over the database (more information on methodology in Chapter 4). The results from the study show that these two new algorithms perform much better than the previously known AIS (R. Agrawal, T. Imielinski, and A. Swami, 1993) and SETM (M. Houtsma and A. Swami, 1993) algorithms. Since the introduction of the Apriori algorithm, it has been considered the most useful and fast algorithm for finding frequent itemsets.

Many improvements have been made on the Apriori algorithm in order to increase its efficiency and effectiveness. (M.J.Zaki, M.Ogihara, S. Parthasarathy, 1996). There are few algorithms developed that are not based on the Apriori,but they still address the issue of speed of Apriori. The following papers (Eu-Hong (Sam) Han, George Karypis, Vipin Kumar, 1999) , (Jong Soo Park, Ming-Syan Chen, Philip S. Yu) propose new algorithms which are not based on the Apriori, but all of them are being compared to Apriori in terms of execution time.

Data Description

The given dataset is a collection of sales records in a large transactional database. The study is based on Indian consumer's buying behavior from a Mall in Spain. The

stores represent products from a local products and brands from other international companies. Using data mining techniques on transactional data leads to the generation of association rules and finding correlations between products in the records. The main concept of association rules is to examine all possible rules between items and turn them into 'if-then' statements.

Association Rule Definition

Let I = { i1,i2,i3,. . . , im} is the set of all items available at the store.By T = {t1,t2,t3, . . . , tn} we define the set of all transactions in the store.Each transaction ti = {i2,i4,i9} contains a subset of items from the whole market basket dataset. An item set is every collection of zero or more items from the transaction database. The number of items that occur in a transaction is called a transaction width.

Let's suppose X is a set of items, e.g. X = {rubber, pencil, sharpener}transaction tj contains an item set X if X is a subset of tj. An association rule can be expressed in the form of X →Y, where X and Y are two disjoint item sets (do not have any items in common).

X is an antecedent and Y is a consequent, in other words, X implies Y. The main concept of association rules is to examine all possible rules between items and turn them into 'if-then' statements. In this case the 'if' part is X or the antecedent, while the 'then' part is Y or the consequent.

Antecedent → consequent [support, confidence]

The antecedent and consequent are often called rule body and rule head accordingly. The generated association rule relates the rule body with the rule head. There are several important criteria of an association rule: the frequency of occurrence, the importance of the relation and the reliability of the rule.

Revised table for functions of association rules (P.D. McNicholas, T.B. Murphy, M. O'Regan)

Function	Definition
Support	S (X→Y) = P (X,Y) and S(X) = P(X)
Confidence	C (X→Y) = P(Y | X)
Expected Confidence	EC (X→Y) = P (Y)
Lift	L(X→Y) = c(X→Y) / P(Y) =
	P (X,Y) / (P(X)P(Y))
Importance	I (X→Y) = log (P(X| Y) / P(Y| not X)

Example 1:

X {[Rubber] + [Pencil]}→Y {[Sharpener]}

Support = 20% Confidence = 50% Lift = 1.5

There are two basic parameters of Association Rule Mining (ARM): support and confidence. (Qiankun Zhao, Sourav S. Bhowmick, 2003) They both measure the strength of an association rule. Since the database is quite large, there is a risk of generating too many unimportant and obvious rules, which may not be of our interest. In that case a common practice is to define thresholds of support and confidence prior to analysis if we want to generate only useful and interesting rules.

Support of an association rule is the percentage of records that contain X U Y to the total number of records in the database. In other words, the support measures how often a rule is applicable to the given dataset. In this measure of strength, quantity is not taken into account. The support count increases by one for each time the item is encountered in a different transaction T from the database D. For example, if a customer buys three packets of pencil in a single transaction, the support count number of [pencil] increases by one.

In other words, the support measures whether an item is present in the transaction or not, ignoring the quantity purchased. If X consists of two items, for example [Pencil] and [Rubber], again the support count number increases by one for every distinct item that is present in the transaction. A high support value means that the rule involves a big part of the database.

Support can be derived from the following formula:

$$\text{Support (XY)} = \frac{\text{Support Count of XY}}{\text{Total Number of Transactions in D}}$$

If the support of X and Y (a set of items) is 10%, it means that X and Y appear together in 10% of the transactions. Retailers will not be interested in items with such low support, as they appear to be purchased together quite rarely. An exception might be when the items of interest are expensive and or generate high profits. Even though such items are rarely purchased, they will be even more profitable if the retailer knows how to exploit the relation between them. In the case with a garment products store we need higher support in order to mine useful and interesting association rules. It is advisable to define

minimum support before the mining process. Specifying the needed minimum support as a threshold prior to analysis generates only itemsets whose supports exceed that given threshold.

However, still there may be some items of interest that are not purchased frequently but give us insightful information. This is the case with expensive and luxury goods in a supermarket, for example. They are not purchased quite often, but the value of the purchase is what matters most. This is why in the aggregation process of the data, more expensive items are rolled up at higher levels of the taxonomy as they do not appear that often in the transactions.

In the given example above, the support of the rule is 20%, which means that the combination of the 3 products occurs in 20% of all transactions.

Confidence of an association rule is defined as the percentage of the number of transactions that contain XUY to the total number of records that contain X. In other words, confidence is a measure of the strength of association rules and is used to determine how frequently items from itemset Y appear in transactions that contain itemset X. Let's suppose we have a rule X →Y. Confidence tells us how likely it is to find Y in a transaction that contains X.

Formula

$$\text{Confidence}(X/Y) = \frac{\text{Support (XY)}}{\text{Support (X)}}$$

In example 1, the confidence is 50%. This means that 50% of all transactions that contain [Pencil] and [rubber] also contain [sharpener] [rubber] occurs in at least 50% of the transactions in which {[pencil] and [sharpener]} occur.

Lift measures the importance of a rule. The lift value is represented as the ratio of the confidence and the expected confidence of a rule. The lift can take over values between zero and infinity. In every association rule we have an antecedent and a consequent, also called rule body and rule head accordingly.

Rule body [Pencil] + [rubber] → Rule head [sharpener]

If the value of the lift is greater than 1 this means that both the rule body and the rule head appear more often together than expected. The occurrence of the rule body positively affects the occurrence of the rule head. The other way around, if the lift value is lower than 1, this means that both the rule body and rule head appear less often together than expected and the occurrence of the rule body negatively affects the occurrence of the rule head. However, if the lift value is near 1, the rule body and rule head appear together as often as expected. (Lift in an association rule)

Lift can be derived from the following formula

$$L(X \rightarrow Y) = c(X \rightarrow Y) / P(Y) = P(X,Y) / (P(X)P(Y))$$

From the given example 1, the lift value is 1.5 which means that the combination of [pencil],[rubber] and [sharpener] is found about 1.5 time more often than expected. However, there is an assumption under which the expected number of occurrences is determined.(See formula for expected confidence in Table 1). The assumption states that the existence of [pencil] and [rubber] in a group does not influence the probability to find [sharpener] in the same group and vice versa.

The Apriori algorithm can be represented in the following steps:

1. Find frequent items and put the to Lk (k=1).
2. Use Lk to generate a collection of candidate itemsets Ck+1 with size (k+1).
3. Scan the database to find which items in Ck+1 are frequent and put them into Lk+1.
4. If Lk+1 is not empty:

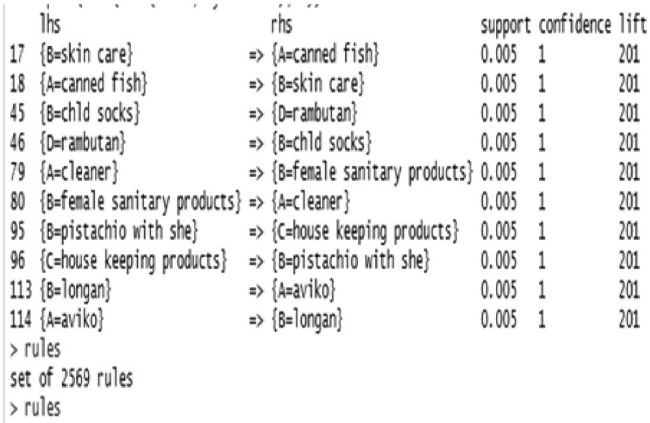
   K:=k+1

   Go to step №2.

```
set of 2569 rules
> options(digits=2)
> rules<-sort(rules, by="confidence", decreasing=TRUE)
>
> inspect(rules[1:10])
   lhs                            rhs                      support confidence lift
2  {D=al ain yog}              => {C=rambutan}             0.005   1          34
3  {D=al ain yog}              => {A=balpamark}            0.005   1          18
4  {D=al ain yog}              => {B=oreo}                 0.005   1          15
5  {D=frozen one man show men} => {C=cheese}               0.005   1          50
6  {D=frozen one man show men} => {B=sadia nuggets}        0.005   1          29
7  {D=frozen one man show men} => {A=olives green}         0.005   1          22
8  {C=rahma}                   => {B=axe oil}              0.005   1          40
9  {C=rahma}                   => {D=dates sukkery loos}   0.005   1          40
10 {C=rahma}                   => {A=kimball}              0.005   1          17
11 {D=balpamark}               => {C=sadia nuggets}        0.005   1          34
>
```

Purchase behavior of boys: output

Purchase behavior of Girls: Output



	lhs		rhs	support	confidence	lift
17	{B=skin care}	=>	{A=canned fish}	0.005	1	201
18	{A=canned fish}	=>	{B=skin care}	0.005	1	201
45	{B=child socks}	=>	{D=rambutan}	0.005	1	201
46	{D=rambutan}	=>	{B=child socks}	0.005	1	201
79	{A=cleaner}	=>	{B=female sanitary products}	0.005	1	201
80	{B=female sanitary products}	=>	{A=cleaner}	0.005	1	201
95	{B=pistachio with she}	=>	{C=house keeping products}	0.005	1	201
96	{C=house keeping products}	=>	{B=pistachio with she}	0.005	1	201
113	{B=longan}	=>	{A=aviko}	0.005	1	201
114	{A=aviko}	=>	{B=longan}	0.005	1	201

```
> rules
set of 2569 rules
> rules
```

Interpretation of the output

The larger the lift the greater the link between the two products. From the two outputs determining what consists of a rule is simply the number of transactions that include both the antecedent and consequent item sets. It is called a support because it measures the degree to which the data support the validity of the rule. In these two outputs we consider only the combination that occurs with higher frequency in the data base. A high value of confidence suggests a strong association rule (in which we are highly confident).A lift ratio is greater than 100 suggests that there is some usefulness to the rule. In other words the level of association between the antecedent and consequent item sets is higher than would be expected if they were independent. The larger the lift ratio, the greater the strength of the association.

Conclusion

In the recent years, more and more retailers are seeking competitive edge through advanced and innovative technology. Market basket analysis is the next step in the retail evolution. Applications of association rule mining are growing rapidly in different sectors – from analysing debit and credit card purchases to fraud detections.

Mining into big data provides managers with a unique window into what is happening with ones business so that they can implement strategies efficiently. Obscure patterns can be discovered using market basket analysis which can help for planning more effective marketing efforts. It can be used not only for cross-sale and up-sale campaigns, but for managing better inventory control and satisfying shoppers' needs. Almost all departments of a company can benefit from a single analysis – not only the high levels of Management but also Store operations, Merchandising and Advertising and Promotion departments.

The market basket problem can be seen as the best example of mining association rules. Discovering association rules has been a well-studied area for the past decade. Building up on previous researches by using established methods for mining association rules allowed for discovering useful information for the retailer

Bibliography

M. Khattak, A. M. Khan, Sungyoung Lee and Young-Koo Lee. (2010). Analyzing Association Rule Mining and Clustering on Sales Day Data with XLMiner and Weka. International Journal of Database Theory and Application Vol. 3, No. 1.

Anderson, C. (2006). The Long Tail: Why the Future of Business is Selling Less of More.

Andreas Mild, Thomas Reutterer. (2003). An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data. Journal of Retailing and Consumer Services vol.10, 123-133.

Andreas Mild, Thomas Reutterer. (2003). An improved collaborative filtering approachfor predicting cross-category purchases based on binary market basket data. Journal of Retailing and Consumer Services, Volume 10, 123-133.

Andrew Ainslie, Peter E. Rossi. (1998). Similarities in Choice Behavior Across Product Categories. Marketing Science, Vol. 17, No. 2, 91-106.

Assunc¸ ˜ao, J. L., & Meyer, R. J. (1993). The rational effect on price promotions on sales and consumption. Management Science, 39 (May), 517–535.

Bari A. Harlam and Leonard M. Lodish. (1995). Modeling Consumers' Choices of Multiple Items. Journal of Marketing Research, Vol. 32, No. 4, 404-418.

Bill Merrilees and Dale Miller. (2001). Superstore interactivity: a new self-service paradigm of retail service? International Journal of Retail & Distribution Management, Vol. 29, Number 8, 379389.

Byung-Do Kim, Kannan Srinivasan, Ronald T. Wilcox. (1999). Identifying Price Sensitive Consumers: The Relative Merits of Demographic vs. Purchase Pattern information. Journal of Retailing, Volume 75(2), 173-193.

Coenen, F. (2011). Data Mining: Past, Present and Future. The Knowledge Engineering Review, Vol. 26:1, 25-29.

D.W. Cheung, A.W. Fu, and J. Han. (1994). Knowledge discovery in databases: a rule based attributeoriented approach. The 8th International Symposium on Methodologies for Intelligent Systems (ISMIS'94),, (pp. 164-173). Charlotte, North Carolina.

David R. Bell and James M. Lattin. (2008). Shopping Behavior and Consumer Preference for Store Price Format: Why "Large Basket". Marketing Science, Vol. 17, No. 1, 66-88.

David R. Bell and YaseminBoztu˘g. (2007). The positive and negative effects of inventory on category purchase: An empirical analysis. Marketing Letters, 18, 1-14.

Eu-Hong (Sam) Han, George Karypis, Vipin Kumar. (1999). Scalable Parallel Data Mining for Association Rules. IEEE Transactions on Knowledge and Data Engineering, vol.20.

Francis J. Mulhern and Robert P. Leone. (1991). Implicit Price Bundling of Retail Products: A Multiproduct Approach to Maximizing Store. Journal of Marketing,Vol. 55, No. 4, 63-76.

Garry J.Russel, Wagner A. Kamakura. (1997). Modeling Multiple Category Brand Preference with Household Basket Data. Journal of Retailing, Volume 73(4), 439-461.

Gary J Russell, Ann Petersen. (2000). Analysis of Cross-Category Dependence in Market Basket Selection. Journal of Retailing, Vol.76(3), 367-392.

Gary J. Russel, Wagner A. Kamakura. (1997). Modeling Multiple Category Brand Preference with Household Basket Data. Journal of Retailing, Volume 73(4), 439-461.

Gregory Piatetsky-Shapiro, William Frawley. (1991). Knowledge Discovery in Databases. AAAI/ MIT Press.

Harald Hruschka, Martin Lukanowicz, Christian Buchta. (1999). Cross-category sales promotion e¤ects. Journal of Retailing and Consumer Services, Volume 6, 99-105.

Jaihak Chung and Vithala R. Rao. (2003). A General Choice Model for Bundles with Multiple-Category Products: Application to Market Segmentation and Optimal Pricing for Bundles. Journal of Marketing Research, Vol. 40, No. 2 , 115-130.

Jong Soo Park, Ming-Syan Chen, Philip S. Yu. (n.d.). Using a Hash-Based Method with Transaction Trimming and Database Scan Reduction for Mining Association Rules. IEEE Transactions on Knowledge and Data Engineering.

Julander, C.-R. (1992). Basket Analysis: A New Way of Analysing Scanner Data. International Journal of Retail and Distribution Management, Volume 20 (7), 10-18.

KatrinDippold, Harald Hruschka. (2010). Variable Selection for Market Basket Analysis. University of Regensburg Working Papers in Business,Economics and Management Information Systems.

Kumar, N. (2009). Kill a brand, keep a customer. Harvard Business Review.

M. Houtsma and A. Swami. (1993). Set Oriented Mining of Association Rules. San Jose, California: IBM Almaden Research Center.

M.J.Zaki, M.Ogihara, S. Parthasarathy. (1996). Parallel Data Mining for Association Rules on SharedMemory Multi-processors. New York: University of Rochester.

❑ ❑ ❑