

Association Analysis

Fan Zhenzhen
Institute of Systems Science
National University of Singapore
E-mail: zhenzhen@nus.edu.sg

© 2016 NUS. The contents contained in this document may not be reproduced in any form or by any means, without the written permission of ISS, NUS, other than for the purpose for which it has been supplied.

Objectives

- To use association analysis as a technique for descriptive analytics
- To be familiar with the situations where association analysis can be applied in business.

Agenda

- Association analysis and its applications
- Link analysis
- Apriori algorithm
- Association rules and evaluation

Modelling

- The process of taking some data and building a simplified description of the processes that might have generated it. The description is often a computer program or mathematical formula.
- A model
 - Is an approximation of the world
 - Expresses some understanding of that world
 - Is expressed in some language (mathematical language, computer language, or modelling language)
- 2 main categories of models and modelling algorithms
 - Descriptive analytics
 - Predictive analytics

Descriptive vs. Prescriptive Analytics

- **Descriptive analytics** – the task of providing a representation of the knowledge discovered without necessarily modelling a specific outcome
 - No specific target variable, therefore: **unsupervised learning**
 - To identify patterns in the data that extend our knowledge and understanding of the world that the data reflects
 - **Cluster analysis** and **association rules**
- **Predictive analytics** – the task of building a model that can be used to predict the occurrence of an event
 - A target variable to be predicted, therefore: **supervised learning**
 - Knowledge extracted from historic data, and the resulting model is applied to new situations
 - Classification and prediction using decision trees, regression, naïve Bayesian network, support vector machines, neural networks, etc.

Descriptive Analytics: Association Analysis

- Has roots in analysis of point-of-sale (POS) transactions, the so-called Market Basket Analysis
 - Determine what products are purchased together or likely to be purchased by the same person
- Common applications (Important and often preferred ways of generating more revenues from the existing customers)
 - Cross-sell - make the purchasers of one product the targets for another
 - Up-sell – target customers likely to upgrade their product or service



Example Applications

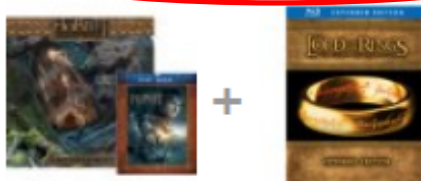
- Items purchased on a credit card (e.g. rental cars, hotel rooms) give insight into the next product the customer may buy
- Optional services bought by telecom customers (call waiting, forwarding, auto-roam etc) show how best to bundle these services
- Banking services used by retail customers (investment services, car loans, home loans, money market accounts etc) show possible cross-sells
- Unusual combinations of insurance claims may indicate fraud
- May find associations between certain combinations of medical treatments and complications in medical patients

In general, when customers do multiple things in close proximity then there is a potential application

Example Applications

- Used widely in targeted advertisement: product recommendation systems
 - Based on items frequently bought in the same transaction (basket or shopping cart)

Frequently Bought Together



Price for both: **\$127.17**

 Add both to Cart

 Add both to Wish List

[Show availability and shipping details](#)

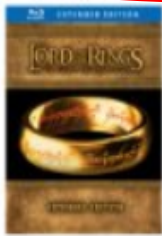
- ☑ **This item:** The Hobbit: An Unexpected Journey Extended Edition with Limited Edition Amazon Exclusive Bilbo/Gollum ... ~ Martin Freeman Blu-ray **\$57.99**
- ☑ The Lord of the Rings: The Motion Picture Trilogy (The Fellowship of the Ring / The Two Towers / The ... ~ Elijah Wood Blu-ray **\$69.18**

Example Applications

- Based on items purchased by the same customer (not necessarily in the same transactions)

Customers Who Bought This Item Also Bought

Page 1 of 25



< The Lord of the Rings: The Motion Picture ...

Elijah Wood

★★★★☆ (6,820)

Blu-ray

\$69.18 ✓Prime



Star Trek Into Darkness (Blu-ray + DVD ...

Chris Pine

★★★★☆ (3,418)

Blu-ray

\$19.85 ✓Prime



Man of Steel (Blu-ray+DVD+UltraViolet ...

Henry Cavill

★★★★☆ (2,143)

Blu-ray

\$19.96 ✓Prime



> The Lord of the Rings: The Return of the King ...

Elijah Wood

★★★★☆ (2,191)

Blu-ray

\$7.99 ✓Prime

Example Applications

- Based on items viewed by the same customer (not necessarily purchased)

Customers Who Viewed This Item Also Viewed

Page 1 of 12

Navigation: < >

Item	Cast	Rating	Price	Prime
The Hobbit: The Battle of Five Armies Extended Edition (BD) [Blu-ray]	Sir Ian McKellen	★★★★☆ 6,277	\$18.95	✓
The Hobbit: The Desolation of Smaug (Extended Edition) (Blu-ray + Digital HD)	Sir Ian McKellen	★★★★☆ 6,269		
The Hobbit: The Battle of the Five Armies (Blu-ray + Downloadable Digital HD)	Sir Ian McKellen	★★★★☆ 6,277	\$13.00	✓
Hobbit: The Motion Picture Trilogy (Extended Edition) [Blu-ray]	Various	★★★★☆ 465	\$63.94	✓
El Hobbit: La Desolación De Smaug	Martin Freeman Ian...	★★★★☆ 28		

Basic MBA

- Requires a list of transactions, each containing a list of items purchased
- E.g. transactions at a convenience store
 - Transaction1: frozen pizza, cola, milk
 - Transaction2: milk, potato chips
 - Transaction3: cola, frozen pizza
 - Transaction4: milk, peanuts
 - Transaction5: cola, peanuts
 - Transaction6: cola, potato chips, peanuts

The Co-occurrence Table

- Cross-tabulate into a table to show how often each possible pair of products were sold together

	Pizza	Milk	Cola	Chips	P/nuts
Pizza	2	1	2	0	0
Milk	1	3	1	1	1
Cola	2	1	4	1	2
Chips	0	1	1	2	1
P/nuts	0	1	2	1	3

Strong Cross-Sell opportunity:

Pizza buyers (2) always also buy cola (2)

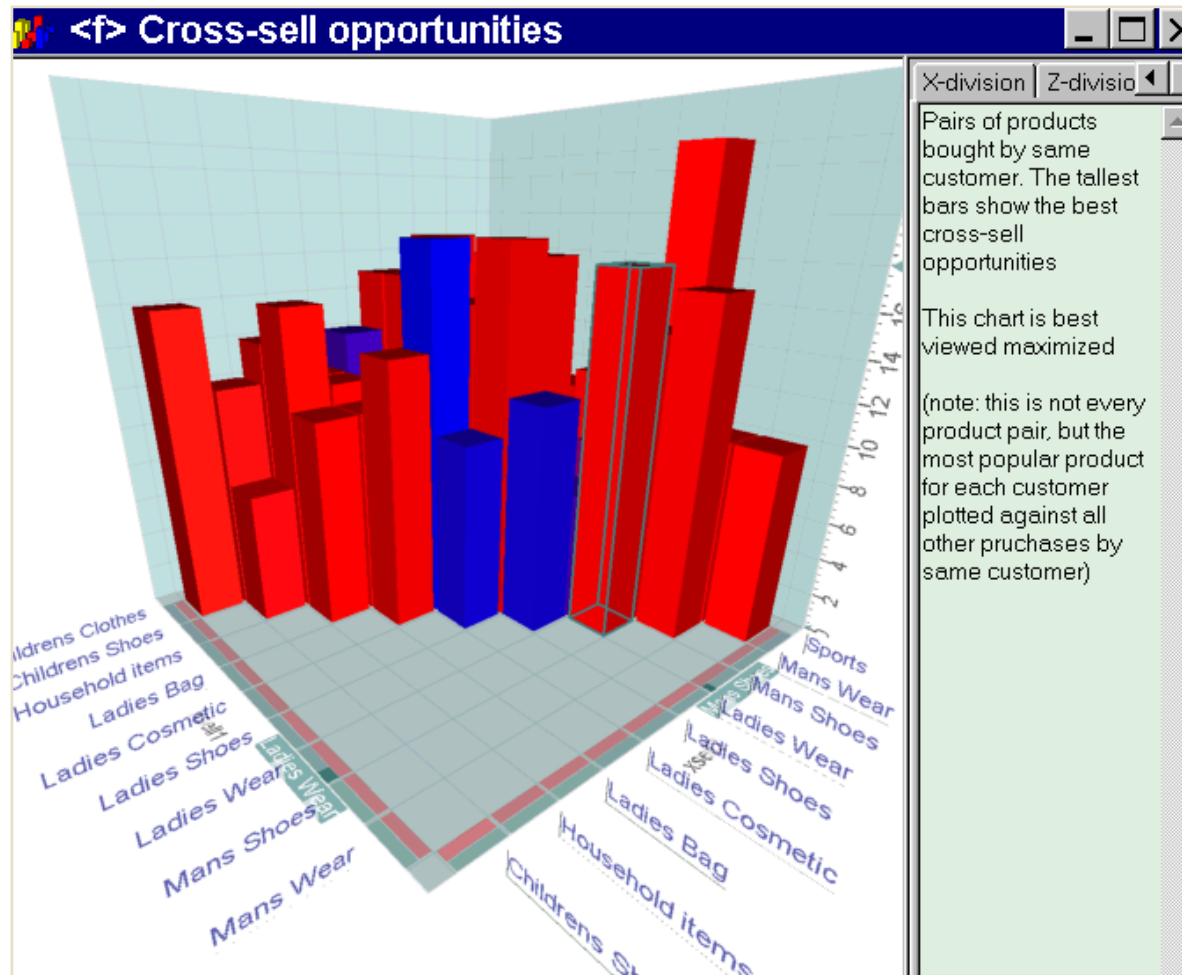
Milk sells well with everything!

Weaker Cross-Sell opportunity:

Peanut buyers (3) nearly always also buy cola (2)

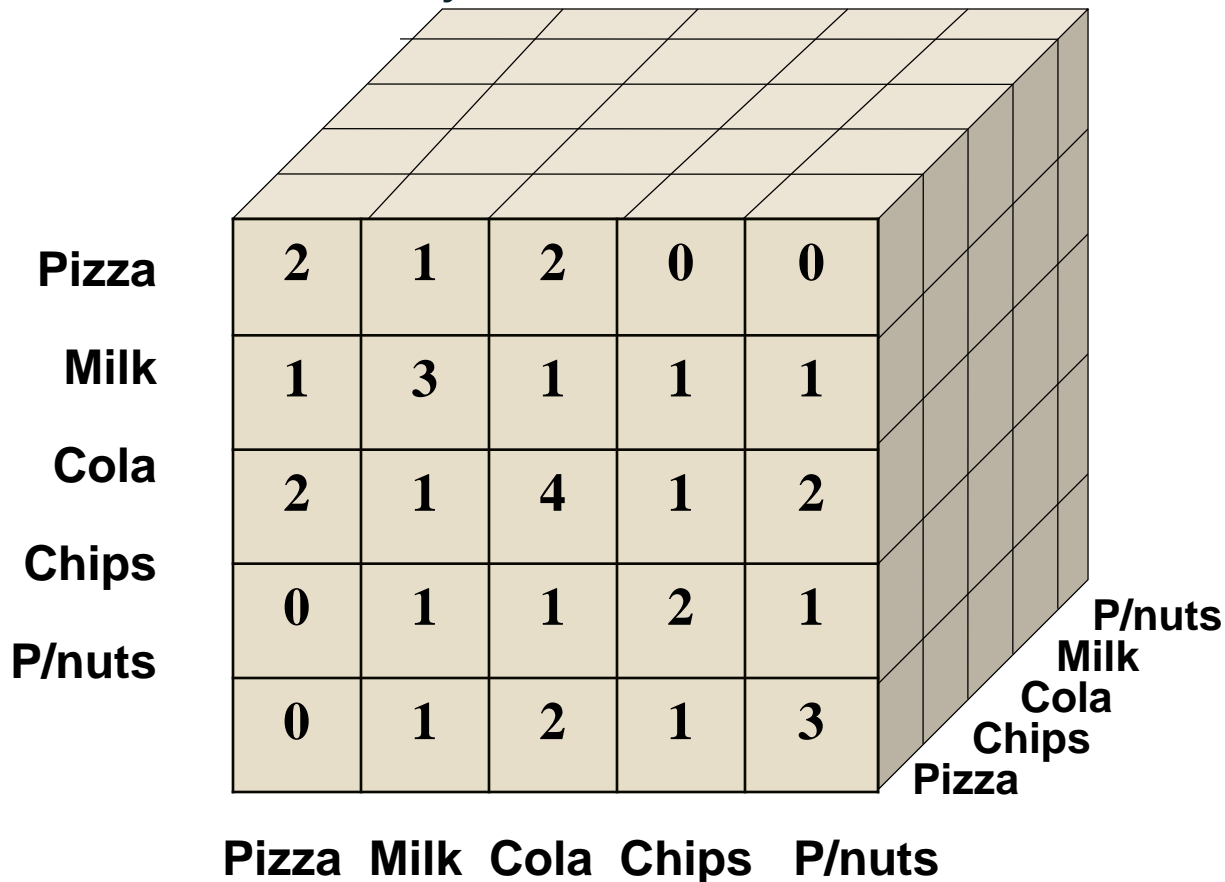
Cola buyers (4) do not always buy pizza (2)

Visualization of Co-occurrence chart



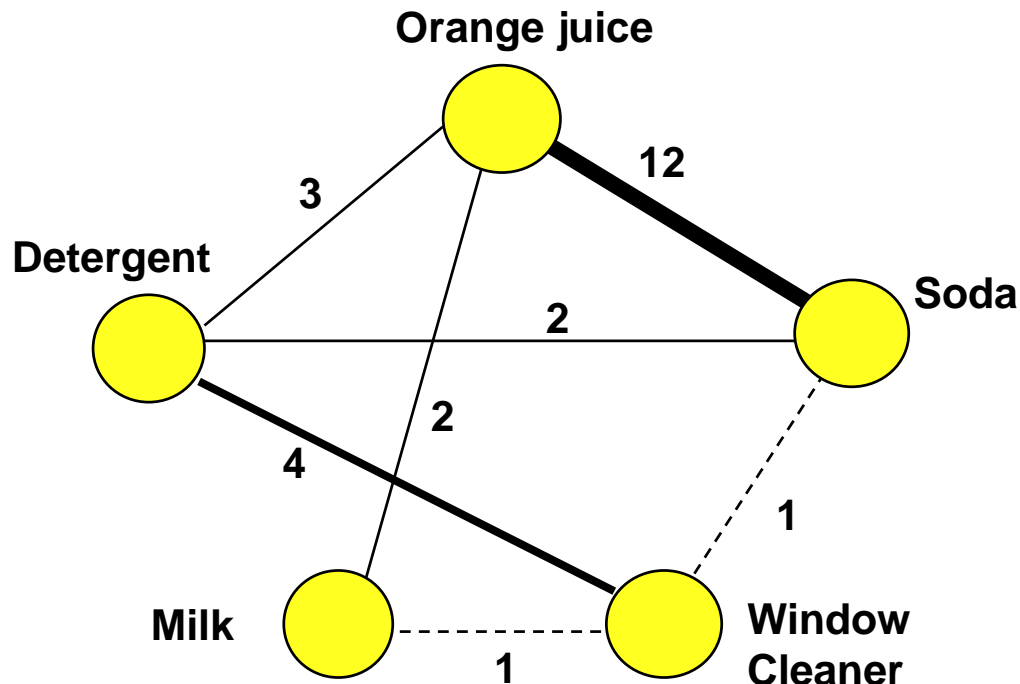
N-way Co-occurrence patterns

- Co-occurrence pairs and triplets can be visualised, higher dimensions cannot be visualised easily

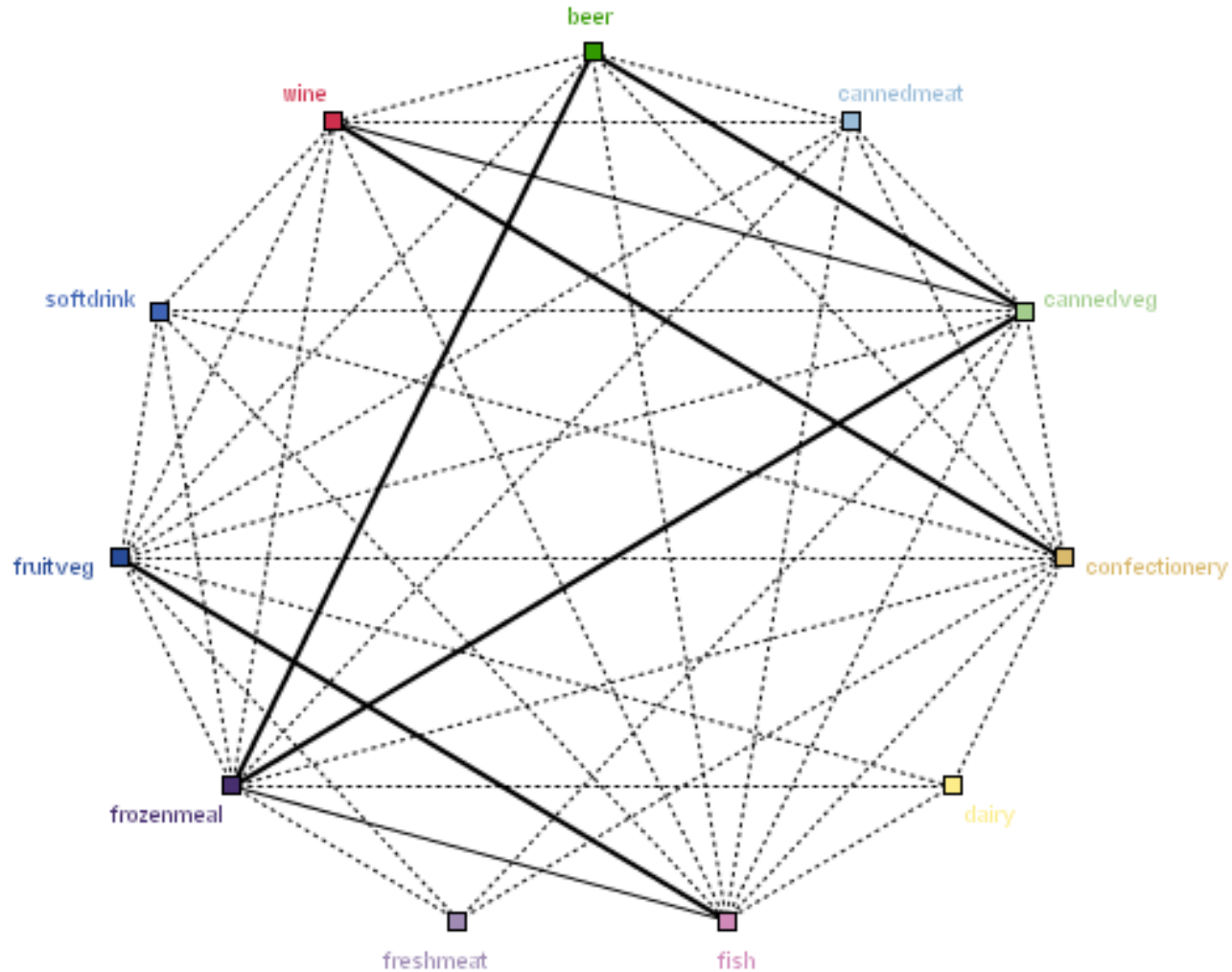


Link Analysis

- Associations can be presented as link graphs
- Nodes represent items; thickness of the joining line indicates the number of times they occurred together



Another Example (IBM SPSS Modeler)



Link Analysis: data format

- Some tools work on raw transactions, while others may require data preprocessing which converts transaction records into summary records for each customer

custID	fruitveg	freshmeat	dairy	cannedveg	beer	wine	softdrink	fish
39808	F	T	T	F	F	F	F	F
67362	F	T	F	F	F	F	F	F
10872	F	F	F	T	T	F	F	T
26748	F	F	T	F	F	T	F	F
91609	F	F	F	F	F	F	F	F
26630	F	T	F	F	F	T	F	T
62995	T	F	F	F	F	F	T	F
38765	F	F	F	F	T	F	F	F
28935	T	F	F	F	F	F	F	F
41792	T	F	F	F	F	F	F	T
59480	T	T	T	T	F	T	F	T
60755	T	F	F	F	F	F	F	T

Tools often provides the transformation function needed, e.g. **Modeler** has a SetToFlag node to do this



Association Rules

- Associations between categorical variables can be found using algorithms and expressed as rules:

<i>If a customer buys Pizza</i>	<i>then he will also buy Cola</i>
<i>LHS</i>	<i>RHS</i>

If Coffee and Milk then Sugar

If BBQ charcoal then Sausages and Steak

Note: Some tools may restrict RHS to one product

- Analysis is based on generating **frequent item sets**
- The algorithm is straightforward. It generally finds the same associations as visual inspection & link analysis, but can take a long time to execute

Association Rule Algorithm (basic)

- Generate all possible rules and examine each according to some criteria.
 - E.g. If there are three items A, B, C, then possible rules are
 - If A and B then C*
 - If A and C then B*
 - If B and C then A*
- Select those whose “goodness” score is above a threshold
- Typical scores (assuming rule “if A then B”)
 - **Support** : probability of getting that combination, also known as *coverage*
 - **Confidence** : $\text{support}(\text{rule}) / \text{support}(\text{LHS})$, also known as *accuracy*
E.g. $\text{confidence}(A \rightarrow B) = \text{support}(A \ \& \ B) / \text{support}(A)$, or $P(B|A)$
 - **Lift** : the increased likelihood in seeing C in a transaction containing A & B
E.g. $\text{confidence}(A \rightarrow B) / \text{support}(B)$, or $P(B|A) / P(B)$

Rule Evaluation Example

- How good are the rules below?
 - **If a customer buys Pizza then they will also buy Cola (R1)**
 - **If a customer buys Peanuts then they will also buy Cola (R2)**
- Data:

Total (100), Pizza (25), Peanut (40), Cola (40)
Pizza & Cola (20), Peanut & Cola (20)
- **Support**
 - Support ~ probability of getting that combination
 - R1 support = 20% (20 trans. out of 100 included pizza & cola)
 - R2 support = 20% (20 trans. out of 100 included peanuts & cola)

Rule Evaluation Example

- **Confidence**

- Confidence = Support combination / Support condition (LHS)
- R1: 80% (20 out of 25 transactions that contain pizza also contain cola)
- R2: 50% (20 out of 40 trans that contain peanuts also contain cola)

- **Lift**

- Lift = confidence (rule) / support (RHS)
- R1: 2 (rule confidence 80% / support of cola 40%)
- R2: 1.25 (rule confidence 50% / support of cola 40%)

Problems with Association Rules

- The basic algorithm is combinatorially explosive

E.g. If 100 products are for sale

Num. items	Num. combinations
1	100
2	4,950
3	161,700
4	3,921,255
5	75,287,520
6	1,192,052,400
8	186,087,894,300

Apriori Algorithm

- Reduces the number of rules to consider by...
 1. Find the **large item sets** from the transaction data
 2. Generate the association rules from the **large item sets**
- **Large item sets, or frequent item sets:** item sets that appear *frequently enough* (threshold parameter) in the data
- Based on the simple observation that all subsets of a frequent item set must also be frequent
 - If $\{milk, bread, cheese\}$ is a frequent item set, so is each of the smaller item sets, $\{milk, bread\}$, $\{milk, cheese\}$, $\{bread, cheese\}$, $\{milk\}$, $\{bread\}$, and $\{cheese\}$
- Significantly reduces search space

Apriori Algorithm

Finding the large item sets

Scan	Candidates	Large item sets
1	{milk}{cola}{pizza} {peanuts}{chips}{mints}	{milk}{cola}{pizza} {peanuts}{chips}
2	{milk cola}{milk pizza} {milk peanuts} {milk chips}{cola pizza} {cola peanuts}{cola chips} {peanuts chips}	{cola peanuts} {cola pizza}
3	{cola peanuts pizza}	

Only consider
item sets with
size > N

Association Rule Examples

Association Rule Set generated by Modeler:

Consequent	Antecedent
frozenmeal	cannedveg
beer	cannedveg
cannedveg	frozenmeal
beer	frozenmeal
frozenmeal	beer
confectionery	wine
wine	confectionery
beer	cannedveg frozenmeal
cannedveg	frozenmeal beer
frozenmeal	cannedveg beer

...

Problems with Association Rules

- Hence can generate a huge number of rules, often trivial and with repetition:

If coffee and milk then sugar

If milk and sugar then coffee

If sugar and coffee then milk

- Define minimum support and minimum confidence for rule pruning/filtering to get “strong” rules
- Analyst must make decisions regarding validity & importance of rules to be accepted (subjective)

Association Rules Examples

Modeler rules
showing *support* &
confidence

Rules have been
sorted by *support*

Consequent	Antecedent	Support %	Confidence %
frozenmeal	cannedveg	30.300	57.100
beer	cannedveg	30.300	55.120
cannedveg	frozenmeal	30.200	57.280
beer	frozenmeal	30.200	56.290
frozenmeal	beer	29.300	58.020
confectionery	wine	28.700	50.170
wine	confectionery	27.600	52.170
beer	cannedveg frozenmeal	17.300	84.390
cannedveg	frozenmeal beer	17.000	85.880
frozenmeal	cannedveg beer	16.700	87.430

...

Association Rules Examples

Sorted by
confidence

Consequent	Antecedent	Support %	Confidence %
cannedveg	freshmeat frozenmeal beer	3.000	96.670
frozenmeal	freshmeat cannedveg beer	3.100	93.550
cannedveg	cannedmeat frozenmeal beer	4.000	90.000
beer	fruitveg cannedveg frozenmeal	4.500	88.890
beer	freshmeat cannedveg frozenmeal	3.300	87.880
frozenmeal	cannedveg beer	16.700	87.430
frozenmeal	fruitveg cannedveg beer	4.600	86.960
beer	dairy cannedveg frozenmeal	2.300	86.960
frozenmeal	dairy cannedveg beer	2.300	86.960
cannedveg	frozenmeal beer	17.000	85.880

...

Association Rule Examples

Association Rule Set generated by SAS Enterprise Miner:



Rule	Confidence(%)	Support(%)	Lift
SVG ==> CKING	87.56	54.17	1.02
CKING ==> SVG	63.15	54.17	1.02
CKING ==> ATM	42.19	36.19	1.10
ATM ==> CKING	94.11	36.19	1.10
SVG ==> ATM	41.53	25.69	1.08
ATM ==> SVG	66.81	25.69	1.08
SVG & CKING ==> ATM	45.88	24.85	1.19
ATM ==> SVG & CKING	64.63	24.85	1.19
SVG & ATM ==> CKING	96.74	24.85	1.13
CKING ==> SVG & ATM	28.97	24.85	1.13
SVG ==> CKING & ATM	40.17	24.85	1.11
CKING & ATM ==> SVG	68.67	24.85	1.11
CKING ==> CD	24.46	20.99	1.00
CD ==> CKING	85.56	20.99	1.00
HMEQLC ==> CKING	100.00	16.47	1.17
CKING ==> HMEQLC	19.20	16.47	1.17
SVG ==> CD	25.40	15.72	1.04
CD ==> SVG	64.08	15.72	1.04
MMDA ==> CKING	89.31	15.58	1.04
CKING ==> MMDA	18.16	15.58	1.04
CKING ==> CCRD	17.32	14.85	1.12
CCRD ==> CKING	95.96	14.85	1.12
SVG ==> CKING & CD	23.04	14.25	1.10

Example of a Misleading “Strong” Rule

- Transactions with respect to the purchase of computer games and videos
 - Total 10,000 transactions
 - 6,000 transactions included computer games
 - 7,500 transactions included videos
 - 4,000 included both
- With min support=30%, min confidence=60%, an association rule is discovered:
 - “buy computer games” => “buy videos” [support=40%, confidence=66%]
- However:
 - Probability of buying videos is actually 75%, even larger than 66%!
 - The association is in fact negative: buying computer games decreases the likelihood of buying videos.

Correlation Analysis Using Lift

- Uses lift to help filter out misleading “strong” association rules
- **Lift** – a simple correlation measure
 - $\text{Lift}(A,B) = P(B|A)/P(B) = P(\{A, B\})/(P(A)P(B))$
 - A is independent of the occurrence of B if $P(\{A,B\})=P(A)P(B)$, ie lift=1
 - Otherwise, A and B are dependent and correlated.
 - Lift >1: positively correlated
 - Lift <1: negatively correlated
- For the rule in the previous slide
 - $\text{lift} = P(\{\text{game, video}\}) / (P(\text{game})P(\text{video})) = 0.40/(0.60 \times 0.75) = 0.89$
=> negative correlation
- Alternative method, the χ^2 measure

Taxonomies and MBA

- At what level of detail should we perform MBA?
- E.g. Should we look for associations between

frozen pizza, chips, peanuts, cola, milk

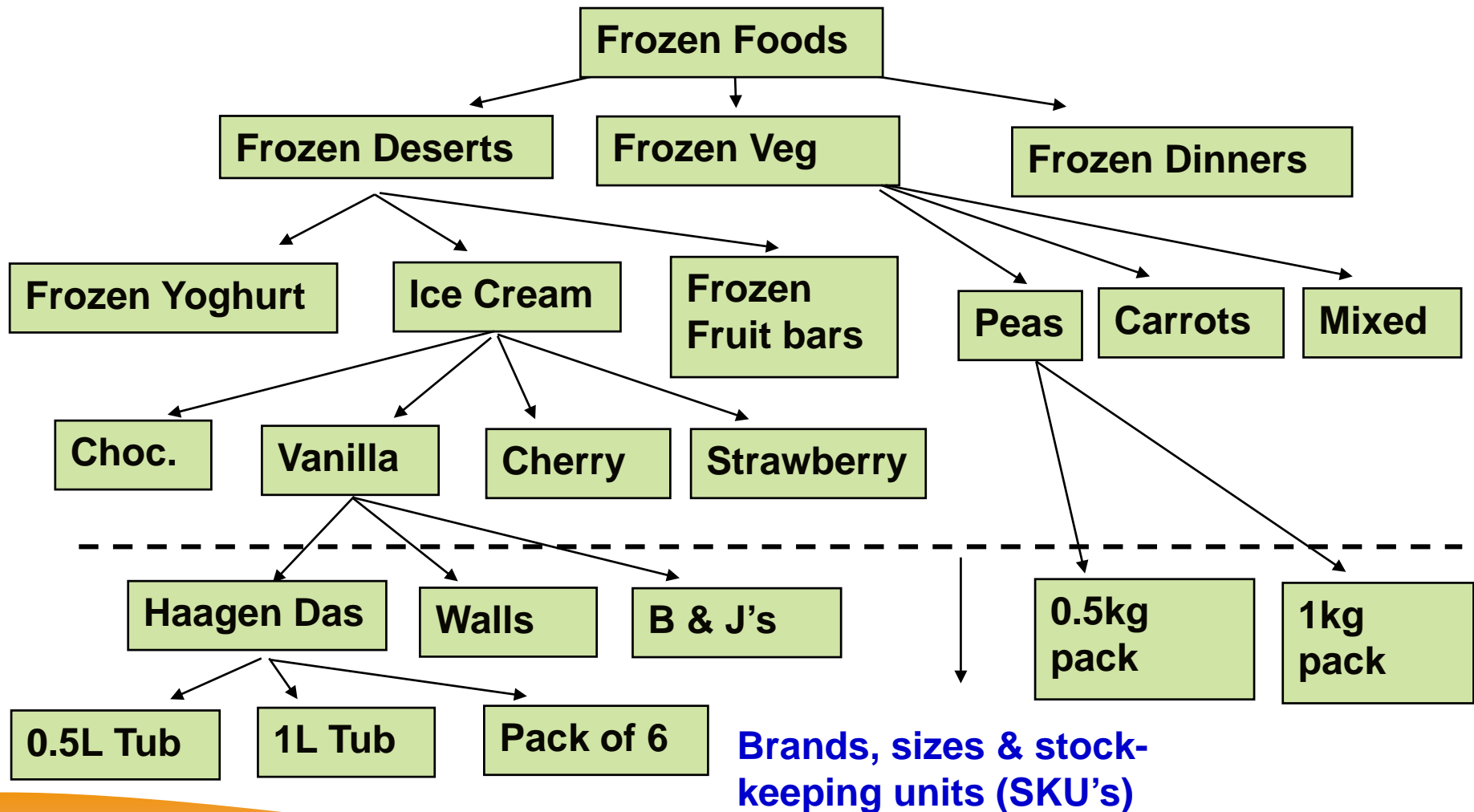
or

*cheese pizza, tuna pizza, salami pizza, vegetable pizza,
chips, peanuts, cola, milk*

Too much detail can
generate an overload of
associations

Too little detail may
yield a non-actionable
result

Products usually form Taxonomies



What level of analysis?

- No need to have all items at the same level
- Level should reflect importance
 - Is brand more important than flavour? Is pack size important?
- Items must occur in approximately the same frequency, otherwise rules are dominated by the common items
 - E.g. 100% of durian cake buyers also buy rice
BUT there is only one durian cake buyer!
- Hence
 - Roll up rare items to higher levels so they are more frequent
 - Break down very frequent items

Virtual Items

- Not real items
- Use to investigate more than which items sell together, extend the association analysis to any categorical variable of interest
- E.g. To find associations between purchased items and new customers
 - A new customer buys a sweater & jacket
 - Enter transaction as:
sweater, jacket, new customer
 - Possible Association Rule:
If new customer and jacket then sweater

Virtual Items

- Other common virtual items
 - Type of promotion
 - Store location (urban, suburban, rural)
 - Season or month, time of day (am, lunch, pm, evening)
 - Payment mode (cash, cheque, credit card)

What about numerical variables?

- Binning is required, partitioning the ranges of quantitative variables into intervals
 - Equal-width binning

The interval size of each bin is the same
 - Equal-frequency binning

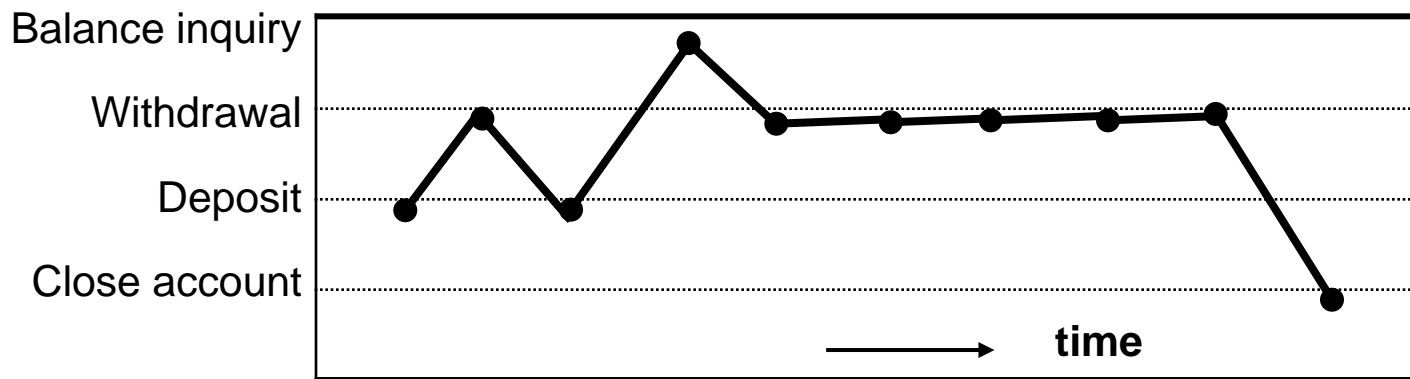
Each bin has approximately the same number of tuples assigned to it
 - Clustering-based binning

Clustering is performed on the variable to group neighboring points (judged based on various distance measures) into the same bin

Sequence Analysis Example

- Can extend MBA to investigate “cause and effect” problems in a sequence dataset (a sequence is an ordering of events)

E.g. What caused the customer to close their account?



- Include the items before and the event of interest as virtual items – possible rule...
 - “Big deposit, withdraw, withdraw, withdraw => close”
- Data must contain identifying information, and time stamp to create sequences or time series

Sequence Analysis Example

E.g. To investigate follow-up treatments administered by different doctors after diagnosis of disease X:

**raw
transactions**

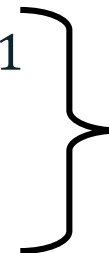


**transactions
after
inserting
virtual items**



Patient ID	Doctor	Sequence	Item
1356	Lim	1	Diagnosis X
5690	Lim	2	Diagnosis X
1356	Lim	3	Prescription 2
7573	Ng	4	Diagnosis X
7573	Ng	5	Prescription 2
5690	Lim	6	Prescription 1
1356	Lim	7	Prescription 1
7573	Ng	8	Prescription 2

- patient 1356, Lim, prescription2, prescription1
- patient 5690, Lim, prescription1
- patient 7573, Ng, prescription2, prescription2



**May suggest
relations between
doctors and
prescription types**

Other Algorithms for Association Rules

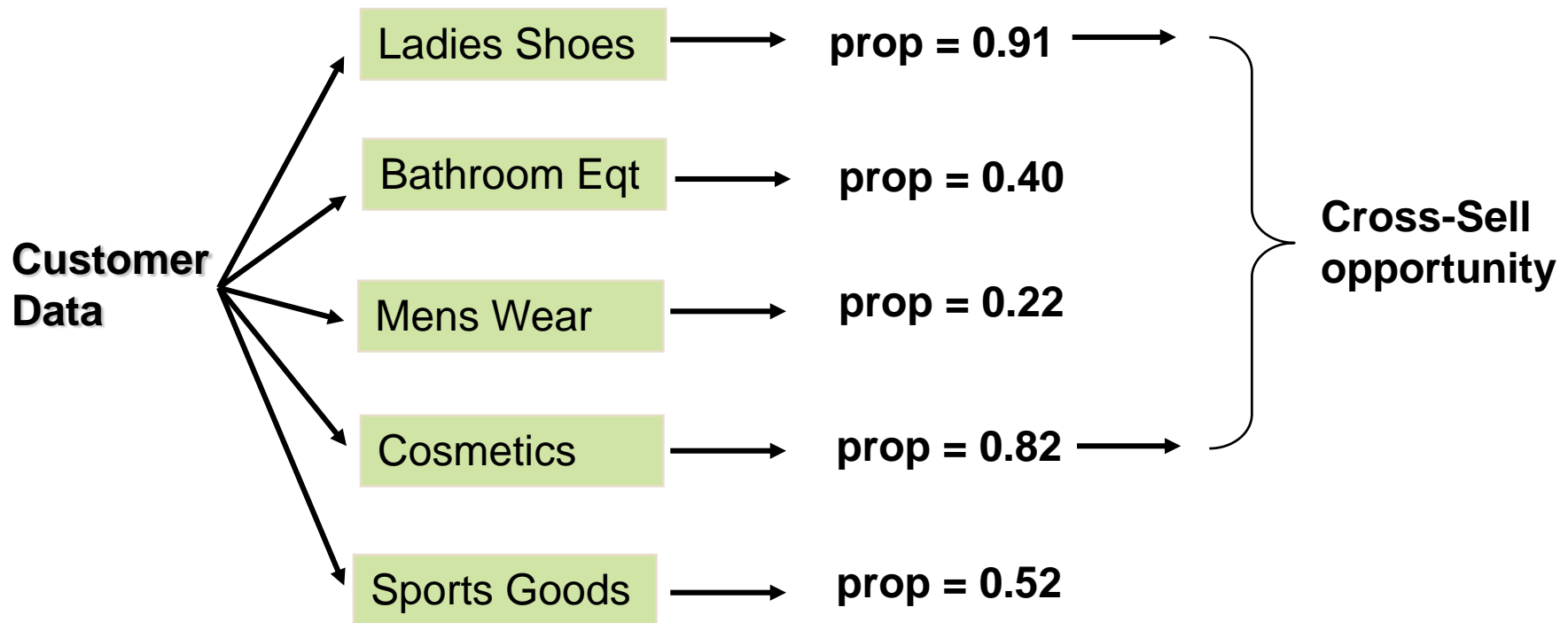
- FP (frequent pattern)-growth algorithm
 - Finding frequent itemsets without candidate generation
 - An FP-tree constructed after two scans of the dataset, then frequent patterns mined from the FP-tree
 - Reducing search cost, more efficient
- ECLAT (Equivalence CLASS Transformation) algorithm
 - Mining vertical data format transformed from transaction data, in which each record indicates an item and a set of transaction IDs containing the item
 - Iteratively generate multi-item sets by, for example, intersecting the TID_sets of every pair of frequent single items

Other Algorithms

- Variations of Apriori
 - Improving efficiency using hash-based techniques, transaction reduction, partitioning, sampling, dynamic itemset counting, etc.
- Algorithms mining
 - Closed frequent itemsets
 - Multilevel association rules
 - Multidimensional association rules
 - Quantitative association rules
- Constraint-based rule mining
- Metarule-guided mining of association rules
- Etc.

Alternative “Cross-Sell” Approaches

- Note that there are other approaches of finding cross-sell opportunities E.g. build a set of models to predict next purchase based on personal data, demographics and past buying history



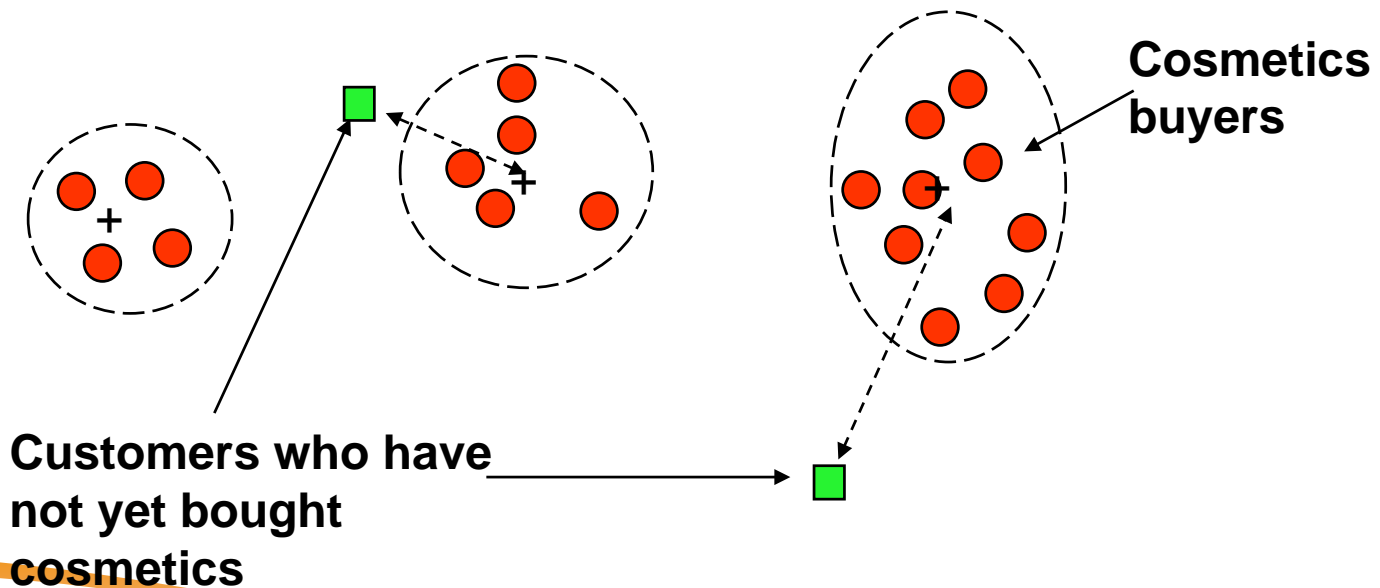
‘Propensity to buy’ Models

“Propensity to buy” models

- More suitable for smaller number of items –otherwise too many models to maintain
- E.g. Cross selling of financial products by a large US bank
 - *Products: brokerage accts, money mkt accts, home loans, savings accts*
 - *Build individual models to predict who is likely to buy each product (each model outputs a propensity score)*
 - *Apply all models to the prospect*
 - *Market the product with the highest score.*

Clustering Approach

- We wish to sell more cosmetics, who should we target?
- Target customers similar to previous cosmetics buyers
 - Cluster the previous cosmetics buyers
 - Use distance to cluster-centre as the similarity measure
 - Market to the closest customers



Summary

- Cross-sell approaches
 - Transaction data ~ co-occurrence matrix
 - Customer-level data
 - link analysis, association rules
 - multiple propensity models, clustering
- Essential to look for associations at appropriate level
 - Use taxonomies to help roll-up & down
- Virtual Items
 - Allow wider applications for MBA techniques