Graduate Certificate in Big Data Analytics

Making Recommendations using Association Mining

Dr. Barry Shepherd
Institute of Systems Science
National University of Singapore
Email: barryshepherd@nus.edu.sg

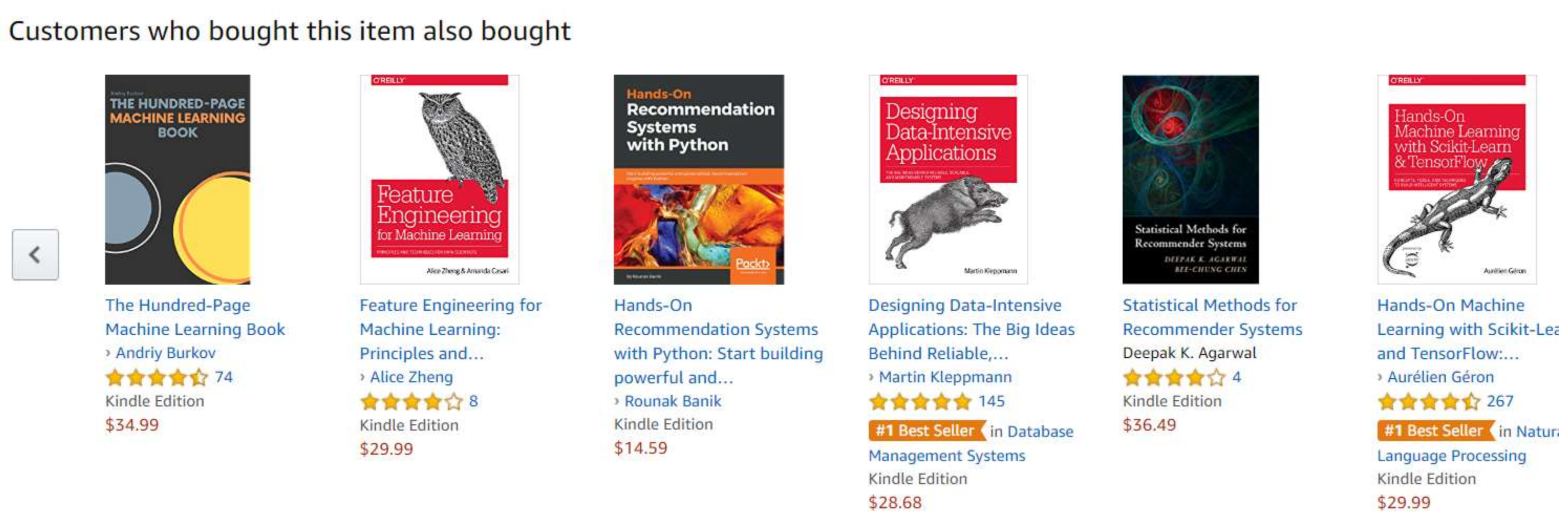
Recommender Systems: Main Approaches

- Simple, non personalised, recommend best selling items, top-rated items, trending items/topics, e.g. as detected via twitter or other social media *(popularity-based)*
- Recommend items that are often bought or viewed at same time as the product you are currently viewing *(market basket analysis)*
- Recommend items similar to those you have already bought or viewed *(content-based filtering)*
- Recommend what people similar to you buy or like *(collaborative filtering)*

Non-personalized recommendations are good for cold-start (new) users

Market-Basket Analysis (MBA)

- The analysis of things that frequently happen together, e.g:
  - Items bought together (in same shopping basket)
  - Items viewed in the same browsing session
- Recommendations made this way are semi-personalized since they are based on broad purchasing trends and they respond to the user's immediate interest (e.g. searching for a particular product) - but don't consider the user's wider likes/dislikes, purchase history etc.
- But…. this is good for situations where the user's intent may be different on every website visit (browsing session) – often called ***session-based*** recommendation



Market Basket Analysis

- Requires a list of transactions
- 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

*Items that occur together (e.g. are in the same basket) are called an **item-set**.*

*If an **item-set** has a large count (is a frequent item-set) then there is a potential association between the items in the item set*

Cooccurrence Counting

- A simple approach to MBA is to cross-tabulate into a table to show how often each possible pair of products were bought together:
- The table (the co-occurrence matrix) is symmetrical since the items in each basket have no temporal ordering. The diagonal shows the total number of times the item was bought.

	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

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

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

Milk sells well with everything

Peanut buyers (3) often buy cola (2)

Association Rule Mining

- Cooccurrence counting is not viable for large datasets and large item-sets
- Algorithms such as Apriori (Agrawal et al, ‘93) use heuristics to reduce the combinatorial size of the search space
  - If an item-set is frequent, then all of its subsets must also be frequent
  - The user specifies minimum rule support and rule confidence
- The found associations are expressed as rules, e.g.

If buy pizza then also buy cola

LHS (left-hand side) RHS (right-hand side)

- In general, association rules can have multiple items in the LHS or RHS

If Coffee and Milk then Sugar

If BBQ charcoal then Sausages and Steak

Note: rules indicate co-occurrence, not causality or a sequence over time

Association Rule Metrics

- Rule Support
  - The proportion of transactions that contain the item set (LHS + RHS items)
- Rule Confidence
  - The proportion of rule firings that are correct predictions
  - Support for combination (LHS & RHS) / Support for condition (LHS)

$$= \frac{\text{\#transactions containing all items in the rule (LHS and RHS)}}{\text{\#transactions containing all items in the rule condition (LHS)}}$$

e.g. consider page-views in a web browsing session as the baskets

session1: home, news, sport
session2: finance, news
session3: fashion, home
session4: news, finance, home
session5: sport, home, finance
session6: fashion, home, news
session7: home, finance, news, sport

home → news
Support = 4/7
Confidence = 4/6

news → home
Support = 4/7
Confidence = 4/5

Note: news→home does not have same *confidence* as home→news

To get intuition consider: whisky->coke (often), but coke->whisky (less)

Association Rule Metrics

$$\text{Rule lift} = \frac{\text{Support (LHS \& RHS)}}{\text{Support (LHS)} * \text{Support (RHS)}} =$$

the ratio of the observed support to that expected if LHS and RHS were independent

If lift = 1 then this implies that the probability of occurrence of X and Y are independent (no association)
If the lift is > 1, then LHS and RHS are dependent on each other (one makes the other more likely)
If the lift is < 1, then one (LHS or RHS) has a negative effect on presence of other

session1: home, news, sport
session2: finance, news
session3: fashion, home
session4: news, finance, home
session5: sport, home, finance
session6: fashion, home, news
session7: home, finance, news, sport

home → news or news → home

Lift = (4/7) / ((6/7)*(5/7))
= 0.57 / 0.61 = 0.93

https://en.wikipedia.org/wiki/Association_rule_learning#Lift

Apriori Algorithm

- **Core concept:** *If an item-set is frequent, then all of its subsets must also be frequent*
- **Example**: Assume 8 transactions (item-sets) and assume support threshold = 3

Itemsets
{1,2,3,4}
{1,2,4}
{1,2}
{2,3,4}
{2,3}
{3,4}
{2,4}

Item	Support
{1}	3
{2}	6
{3}	4
{4}	5

Count the support for all individual items (the 1-itemsets). Ignore those with support < 3

Item	Support
{1,2}	3
{1,3}	1
{1,4}	2
{2,3}	3
{2,4}	4
{3,4}	3

Examine only the frequent 1-itemsets.
Count support for pairs of items (the 2-itemsets)

Item	Support
{2,3,4}	2

Examine only the frequent 2-itemsets.
Count support for triplets of items (the 3-itemsets)

Efficient data structures (e.g. tree-based search) are used to implement

Association Mining 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

Issues with Association Rules

- 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)

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

e.g. rules sorted by support

Example: MSNBC Website Mining

- Data from the Pagview log of MSNBC.com (a news site) on 28 Sep'99
  - Raw data ~ Datetime, URL, cookieID, IPaddress, UserAgent(browser) etc...

- Preprocessing
  - **Page view categorization** – the pages viewed (URLs) were first converted into topic categories
  - **Grouping by unique user** - each data row shows the sequence of pageview categories for one user on that day

Codes for the msnbc.com page categories

category	code	category	code	category	code
frontpage	1	misc	7	summary	13
news	2	weather	8	bbs	14
tech	3	health	9	travel	15
local	4	living	10	msn-news	16
opinion	5	business	11	msn-sport	17
On-air	6	sports	12		

```
% Sequences:

6
1 1
6
6 7 7 7 6 6 8 8 8 8
6 9 4 4 4 10 3 10 5 10 4 4 4
1 1 1 11 1 1 1
12 12
```

Number of users: 989,818
Average number of visits per user: 5.7
No. of URLs per category: 10 to 5000

Example MSNBC Rules*

- Association Rule Examples

on-air & business & sports & bbs	-->	frontpage	86.22%
news & tech & misc & bbs	-->	frontpage	86.18%
on-air & misc & business & sports	-->	frontpage	86.16%
tech & misc & travel	-->	on-air	86.09%
tech & living & business & sports	-->	frontpage	86.08%
news & living & sports & bbs	-->	frontpage	85.99%
misc & business & sports	-->	frontpage	85.79%

- Sequence Rules Examples

on-air → news	1.51%
news → frontpage → news	1.49%
local → news	1.46%
frontpage → frontpage → business	1.35%
news → sports	1.33%
news → bbs	1.23%
health → local	1.16%

Are these rules likely to be useful?

*(*Frequent Pattern Mining in Web Log Data, Ivancsy, Vajk, 2006. Using their own association and sequence finding algorithms)*

YouTube Video Recommendations

https://www.inf.unibz.it/~ricci/ISR/papers/p293-davidson.pdf (2010)

Step1: Use *Association Mining* to generate a *seed-set* of candidate videos:

- For each video pair sequence, count how often they were viewed in same session (e.g. within 24hours)
- Given a seed video, select the top N related videos ranked by their normalized co-occurrence counts. Impose a minimum score threshold.
- To personalise, add videos that user watched, liked, rated, or added to playlists.
- To diversify, expand *seed set* using graph traversal: add neighboring videos

Step2: Score & rank the candidates using:

- User independent: overall rating of a video, #times watched, …
- User specific: view count and time of watch of each seed video (if previously watched), ..
- Diversification: limit #recommendations from a single seed video, or from same channel, …

What Level of Detail to Recommend?

- At what level of detail should we make the recommendations?
- E.g. Should we look for associations between:

frozen pizza, chips, cola, milk

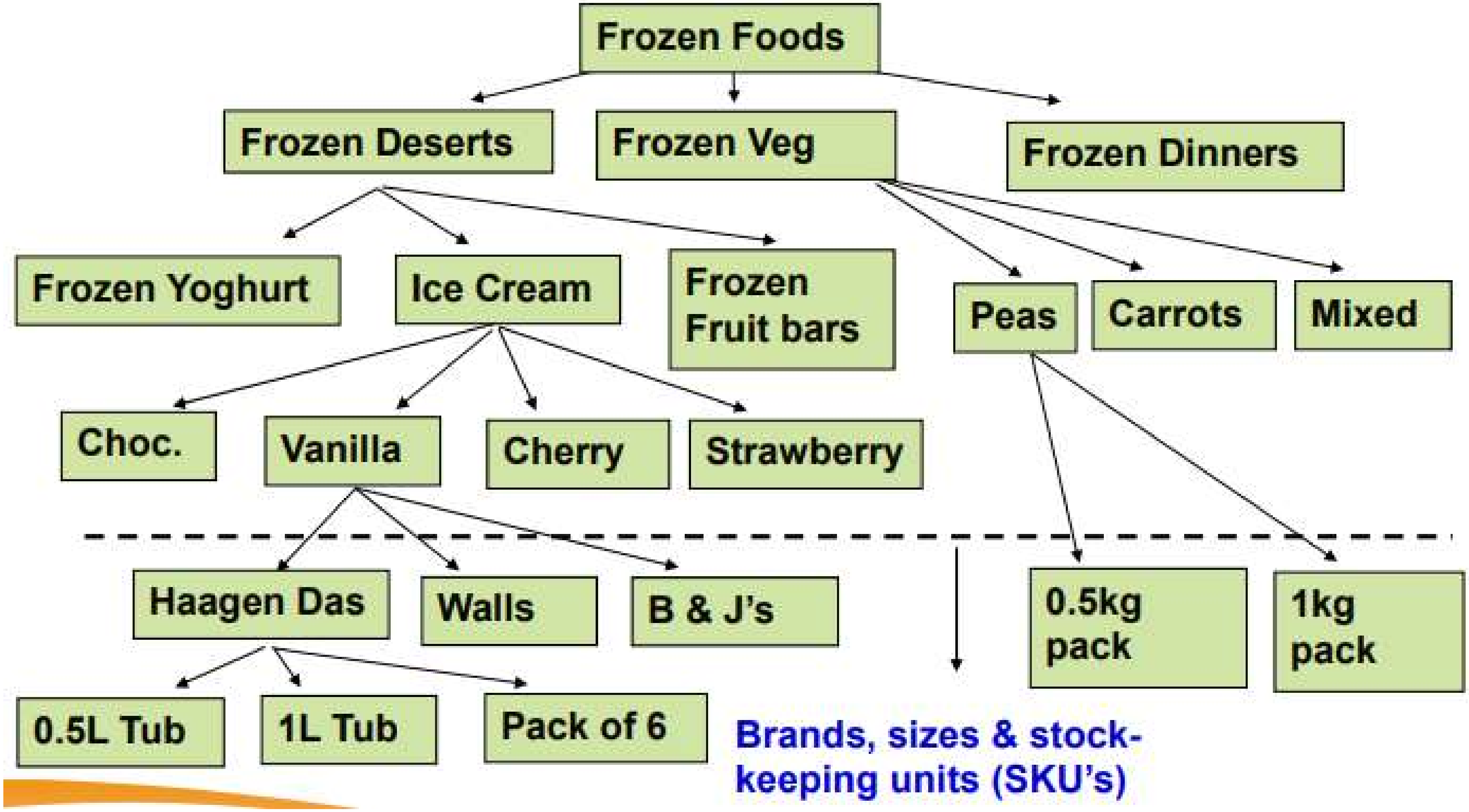
Or:

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

Too much detail can generate an overload of associations

Too little detail may yield obvious and non-actionable results

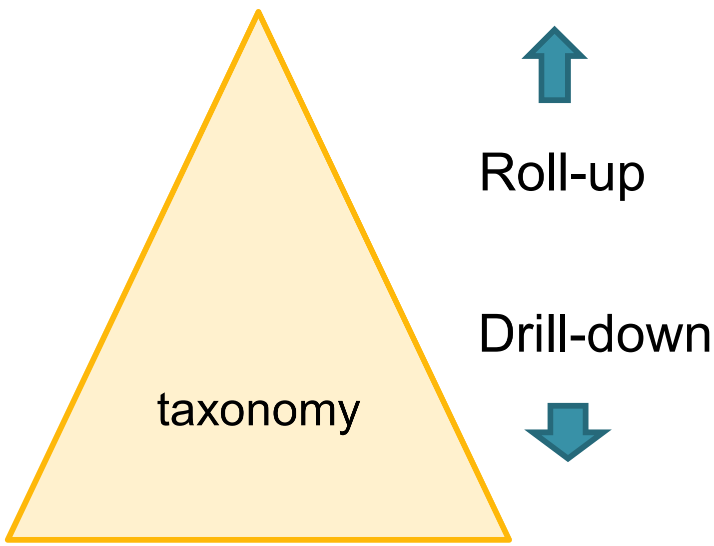
Products often form Taxonomies



- Note that the taxonomy is often not unique

Recommending with Taxonomies

- 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 should occur in approximately the same frequency, otherwise rules are dominated by the common items
  - E.g. rice is bought very frequently and durian cakes very infrequently, hence the rule: *if durian cake then rice* is likely to be found, but is not useful
- Hence …..
  - Roll-up rare items to higher levels so they are more frequent
    - e.g. salami pizza, tuna pizza, veg.pizza etc -> pizza
  - Break-down (drill-down into) very frequent items
    - e.g. pizza -> salami pizza, tuna pizza, veg.pizza etc



Virtual Items

- Expand the scope of association mining from items / products to any categorical variable of interest e.g.
  - Type of promotion
  - Store location (urban, suburban, rural)
  - Season or month, time of day (am, lunch, pm, evening)
  - Payment mode (cash, cheque, credit card)
  - Gender of customer (male, female)
- Can add any relevant information as an ‘item’ into the basket
- E.g. To find associations between purchased items and new customers
  - Enter transaction as: {sweater, jacket, new customer}
  - Possible Association Rule: *If new customer and jacket then sweater*
- Can include numerical concepts into the baskets by first binning them
  - E.g. Age -> young-age, middle-age, senior-age
  - Possible Association Rule: *If outdoor jacket and middle-age then hiking boots*

Sequence Mining

- Similar to Association Mining, but the order items are put into the basket is important
  - Can be a sequence of events over time, DNA sequences, word sequences, ….
- Good for Session-Based Recommender Systems / Clickstream Mining

Web-site browsing is often done by anonymous users identifiable only via a cookie ID. Cookies don’t last long hence historical user data is usually unavailable - we only have their browsing behaviour in the current session

The previous example….

session1: home, news, sport, finance
session2: finance, news
session3: fashion, home, sport
session4: news, finance, home
session5: sport, home, finance
session6: fashion, home, news, news, sport
session7: home, finance, news, finance, sport

Frequent sub-sequence:
home-> news -> sport

Methods For Sequence Mining

- Apriori-based Approaches
  - GSP (Generalised Sequential Pattern, *Agrawal and Srikant'96*)
  - SPADE (Sequential Pattern Discovery using Equivalence Classes, *Zaki'01*)
- Pattern-Growth-based Approaches
  - FreeSpan (Frequent Pattern-projected Sequential Pattern Mining) *(Han et al.@KDD'00)*
  - PrefixSpan (Prefix-projected Sequential Pattern Mining) *(Pei, et al.@ICDE'01)*
- More Advanced Sequence/Session Mining
  - Deep learning approaches leveraging NLP methods – e.g. treat a sequence of clicks as a sentence ….. *more on this in day3*

The cat sat on the → mat

Workshop1: Association Mining and Recommender Systems

- Building association rules
- Rule execution and testing
- Comparison with predictive modelling methods (supervised learning)