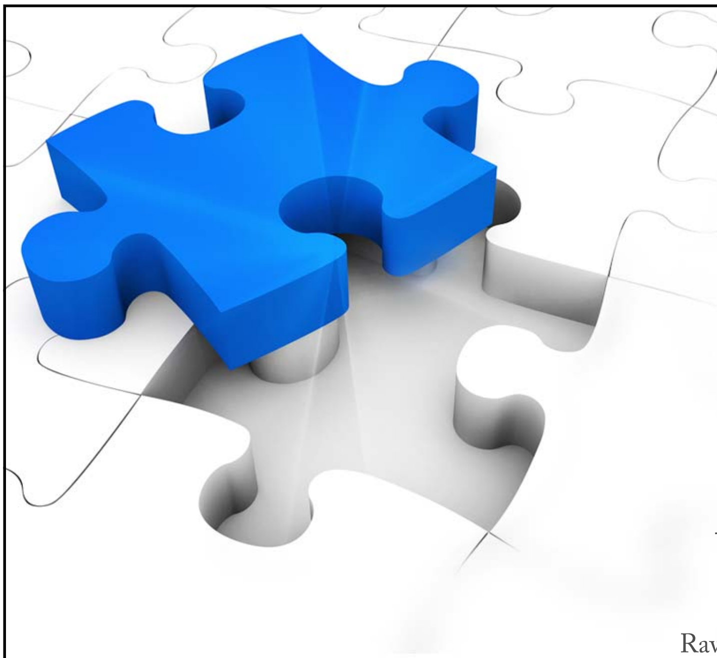


Lecture 09:
Association Analysis

Rawls Profess of MIS
Jaeki Song, Ph.D.

Association

Affinity Analysis and Market Basket Analysis

- Affinity analysis
  - the study of attributes that "go together"
  - known as "market basket analysis"
    - seek to uncover associations among these attributes
- Association rules
  - "if *antecedent*, the *consequent*"

Affinity Examples

- Proportion of subscribers to cell phone plan that respond positively to an offer of service upgrade
- Proportion of children whose parents read to them who are themselves good readers
- Predicted degradation in telecommunications networks
- Finding which items in a supermarket are purchased together and which are never purchased together
- Proportion of cases in which new drug will exhibit dangerous side effects

Data Representation for Market Basket Analysis

- Transactions made at the roadside vegetable stand

Transition	Items Purchased
1	Broccoli, green peppers, corn
2	Asparagus, squash, corn
3	Corn, tomatoes, beans, squash
4	Green papers, corn, tomatoes, beans
5	Beans, asparagus, broccoli
6	Squash, asparagus, beans, tomatoes,
7	Tomatoes, corn
8	Broccoli, tomatoes, green peppers
9	Squash, asparagus, beans
10	Beans, corn
11	Green peppers, broccoli, beans, squash
12	Asparagus, beans, squash
13	Squash, corn, asparagus, beans
14	Squash, corn, asparagus, beans
15	Corn, green peppers, tomatoes, beans, broccoli

- transactional data format for the roadside vegetable stand data

Affinity Analysis

Transactional Data Format Excerpt of first 4 rows:

Transaction ID	Items
1	Broccoli
1	Green peppers
1	Corn
2	Asparagus

Tabular Data Format Excerpt of first 4 rows:

Transaction	Asparagus	Beans	Broccoli	Corn	Green Peppers	Squash	Tomatoes
1	0	0	1	1	1	0	0
2	1	0	0	1	0	1	0
3	0	1	0	1	0	1	1
4	0	1	0	1	1	0	1

Notations

- D: the set of transactions, where each transaction represents a set of items contained in I
  - A: particular set of items (e.g., bean, squash, etc.)
  - B: Another set of items (e.g. asparagus)
- Association rule takes the form
  - if *A*, then *B* (A => B)
    - A and B are mutually exclusive
    - A (left-hand side) and B(right-hand side)

Notations

- An *itemset* is set of items contained in *I* a *k-itemset* is an itemset containing *k* items
- An *itemset frequency* is number of transactions containing itemset
- A *frequent itemset* is an itemset that occurs at least a minimum, ϕ times
- The set of *frequent k*-itemsets is denoted F_k

Association Rule

- Support
  - a particular association rule A=> B is the proportion of transactions in D that contains both A and B
  - $P(A \cap B) = \frac{\#\ of\ transactions\ containing\ both\ A\ and\ B}{total\ \#\ of\ transactions}$
- Confidence
  - A measure of the accuracy of eh rule, as determined by the percentage of transactions in D containing A that also contain B
  - $P(B|A) = \frac{P(A \cap B)}{P(A)}$

$$= \frac{\#\ of\ transactions\ containing\ both\ A\ and\ B}{\#\ of\ transactions\ containing\ A}$$

Association Rule

- Support is the proportion of transactions that contain both A and B
- Confidence is a measure of the accuracy of the rule as determined by the percentage of transactions in D containing A that also contain B
  - strong rules are those that meet or surpass minimum support and confidence criteria

Association Rule

- Mining association rules
  - Find all frequent itemsets
    - find all itemsets with frequency ≥ ϕ
  - From the frequent itemsets, generate association rules satisfying the minimum support and confidence conditions.
- A prior property
  - If an itemset Z is not frequent then for any item A, Z ∪ A will not be frequent.

Association Rule

- Lift
  - $Lift = \frac{Rule\ confidence}{Prior\ proportion\ of\ the\ consequent} = \frac{P(A \cap B)}{P(A)P(B)}$
    - $\text{lift}(A \Rightarrow B) > 1$
    - $\text{lift}(A \Rightarrow B) < 1$
- Analysts prefer rules that have either high support or high confidence, and usually both
  - Rules with lift values different from 1 will be more interesting and useful than those with lift values near 1

Association Rule

- Example

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

$\text{support}(Diaper \Rightarrow Beer) = {}^{3}/_{5}$
$\text{confidence}(Diaper \Rightarrow Beer) = {}^{3}/_{4}$
$\text{lift}(Diaper \Rightarrow Beer) = {}^{1}/_{4}$

Example

- Categorical data → binarization

CUST_ID	GENDER	AGE	CHILD_PRD_YN	MOBILE_APP_USE	RE_ORDER
1	FEMALE	23	NO	YES	YES
2	MALE	28	NO	YES	NO
3	FEMALE	42	NO	NO	NO
4	FEMALE	34	YES	YES	YES
5	MALE	45	NO	NO	NO
6	FEMALE	36	YES	YES	YES

CUST_ID	GENDER = MALE	GENDER = FEMALE	AGE =20	AGE =30	AGE =40	CHILD_PRD_YN =YES	CHILD_PRD_YN =NO	MOBILE_APP_USE =YES	MOBILE_APP_USE =NO	RE_ORDER =YES	RE_ORDER =NO
1	0	1	1	0	0	0	1	1	0	1	0
2	1	0	1	0	0	0	1	1	0	0	1
3	0	1	0	0	1	0	1	0	1	0	1
4	0	1	0	1	0	1	0	1	0	1	0
5	1	0	0	0	1	0	1	0	1	0	1
6	0	1	0	1	0	1	0	1	0	1	0

Example

```
##      lhs                         rhs             support confidence lift
## 1  {child_prd_yn=YES}         => {re_order=YES} 0.3333333          1    2
## 2  {child_prd_yn=YES,
##     mobile_app_use=YES}       => {re_order=YES} 0.3333333          1    2
## 3  {gender=FEMALE,
##     child_prd_yn=YES}         => {re_order=YES} 0.3333333          1    2
## 4  {child_prd_yn=YES,
##     age_cd=age_20}            => {re_order=YES} 0.3333333          1    2
## 5  {gender=FEMALE,
##     mobile_app_use=YES}       => {re_order=YES} 0.5000000          1    2
## 6  {gender=FEMALE,
##     child_prd_yn=YES,
##     mobile_app_use=YES}       => {re_order=YES} 0.3333333          1    2
## 7  {child_prd_yn=YES,
##     mobile_app_use=YES,
##     age_cd=age_20}            => {re_order=YES} 0.3333333          1    2
## 8  {gender=FEMALE,
##     child_prd_yn=YES,
##     age_cd=age_20}            => {re_order=YES} 0.3333333          1    2
## 9  {gender=FEMALE,
##     mobile_app_use=YES,
##     age_cd=age_20}            => {re_order=YES} 0.5000000          1    2
## 10 {gender=FEMALE,
##     child_prd_yn=YES,
##     mobile_app_use=YES,
##     age_cd=age_20}            => {re_order=YES} 0.3333333          1    2
```

Example

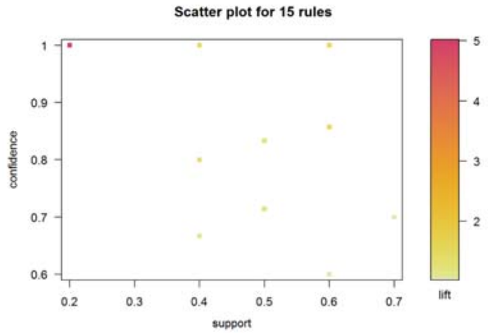
- DVD data

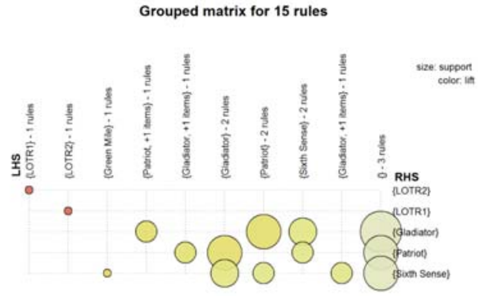
id	Item
1	Sixth Sense
1	LOTR1
1	Harry Potter1
1	Green Mile
1	LOTR2
2	Gladiator
2	Patriot
2	Braveheart
3	LOTR1
3	LOTR2
4	Gladiator
4	Patriot
4	Sixth Sense
5	Gladiator

```
> inspect(rules)
   lhs                  rhs       support   confidence lift
1  {Alista}          => {Nami}    0.1111111 1          9.0
2  {Nami}            => {Alista}  0.1111111 1          9.0
3  {Lammus}          => {Malpike} 0.1111111 1          4.5
4  {Sona}            => {Kaytlne} 0.1111111 1          3.6
5  {Nasus}           => {Zaira}   0.1111111 1          4.5
6  {Nasus}           => {Amumu}   0.1111111 1          3.6
7  {Kaytlne,Shinzao} => {Amumu}   0.1111111 1          3.6
8  {Amumu,Shinzao}   => {Kaytlne} 0.1111111 1          3.6
9  {Nasus,Zaira}     => {Amumu}   0.1111111 1          3.6
10 {Amumu,Nasus}     => {Zaira}   0.1111111 1          4.5
11 {Amumu,Zaira}     => {Nasus}   0.1111111 1          9.0
12 {Kaytlne,Malpike} => {Amumu}   0.1111111 1          3.6
13 {Amumu,Malpike}   => {Kaytlne} 0.1111111 1          3.6
```

Example

- DVD Example





Example

- Grocery data

ID	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9
1	yogurt	water	pastry	shopping b	tropical fru	soda	sausage		
2	bread	tropical fru	pastry	vegitables	soda	sausage	milk	shopping bag	
3	bread	pastry	tropical fru	water	yogurt	vegitables	sausage	shopping bag	
4	shopping b	water	pastry	bread	sausage	vegitables			
5	vegitables	yogurt	soda	sausage	water	shopping bag			
6	soda	pastry	tropical fru	bread	shopping b	sausage			
7	sausage	tropical fru	bread						
8	vegitables	bread	tropical fru	milk	yogurt	water	sausage	pastry	shopping bag
9	water	pastry	bread	vegitables	shopping bag				
10	vegitables	sausage	pastry	bread	yogurt	shopping bag			
11	tropical fru	bread	sausage	water	yogurt	soda	shopping bag		
12	vegitables	sausage	tropical fru	pastry	bread	shopping bag			
13	vegitables	yogurt	sausage						
14	tropical fru	soda	vegitables	pastry	shopping b	water	bread		
15	tropical fru	soda	bread	shopping b	sausage	water	vegitables	pastry	
16	tropical fru	soda	water	sausage	shopping bag				
17	soda	bread	milk	vegitables	shopping bag				
18	tropical fru	pastry	water	shopping b	bread				
19	tropical fru	shoppng ba	soda	water	vegitables	pastry	yogurt	milk	
20	soda	pastry	shopping b	vegitables	sausage	bread	tropical fru	water	

Example

- Read transaction data
  - read.transactions( )

```
inspect(rules)
```

```
514 {bread,pastry,sausage,tropical fruit,vegitables,water}         => {shopping bag}   0.1904762 1.0000000  1.400000
515 {pastry,sausage,shopping bag,tropical fruit,vegitables,water} => {bread}          0.1904762 1.0000000  1.400000
516 {bread,pastry,shopping bag,tropical fruit,vegitables,water}   => {sausage}        0.1904762 0.8000000  1.120000
517 {bread,sausage,shopping bag,tropical fruit,vegitables,water}  => {pastry}         0.1904762 1.0000000  1.500000
518 {bread,pastry,sausage,shopping bag,vegitables,water}          => {tropical fruit} 0.1904762 1.0000000  1.500000
519 {bread,pastry,sausage,shopping bag,tropical fruit,water}      => {vegitables}     0.1904762 1.0000000  1.500000
```

Example

- Parameter adjustment
  - minlen = 2, supp= 0.3, conf= 0.9
  - Sort the outputs
    - quality( ), round ( ), and sort( )

```
quality(rules)<-round(quality(rules), digits=3)
rules.sorted<- sort(rules, by = 'confidence')
inspect(rules.sorted)
```

```
##    lhs                  rhs              support confidence  lift
## 1  {pastry,
##     soda}            => {tropical fruit}   0.333      1.000 1.500
## 2  {bread,
##     soda}            => {shopping bag}     0.333      1.000 1.400
## 3  {tropical fruit,
##     vegitables}      => {pastry}           0.381      1.000 1.500
## 4  {pastry,
##     shopping bag}    => {bread}            0.524      1.000 1.400
```