

Collaborative Filtering (II)

Qiang Yang
Simon Fraser University
Thanks: Jiawei Han

4/4/01 collaborative filtering 1

Association Rules

- Find rules of the form:
 - "Users who buy items i, j , also buy k "
 - Item(i) and Item(j) \rightarrow Item(k)
 - Confidence
 - Support
- Association Rules can then be applied to a new user X
 - If X buys i, j , then predict that X will buy k

4/4/01 collaborative filtering 2

Example

- Which movie would Sammy watch next?
 - Ratings 1--5

	Titles				
	Starship Trooper (A)	Sleepless in Seattle (R)	MI-2 (A)	Matrix (A)	Titanic (R)
Sammy	3	4	3	?	?
Beatrice	3	4	3	1	1
Dylan	3	4	3	3	4
Mathew	4	2	3	4	5
Gum-Fat	4	3	4	4	4
Basil	5	1	5	?	?

- Matrix: 3, Titanic: 14/4; Recommend Titanic!

4/4/01 collaborative filtering 3

What Is an Association Rule?

- Given
 - A database of customer transactions
 - Each transaction is a list of items (purchased by a customer in a visit)
- Find all rules that correlate the presence of one set of items with that of another set of items
 - Example: 98% of people who purchase tires and auto accessories also get automotive services done
 - Any number of items in the consequent/antecedent of rule
 - Possible to specify constraints on rules (e.g., find only rules involving Home Laundry Appliances).

4/4/01 collaborative filtering 4

Application Examples

- Market Basket Analysis
 - * \Rightarrow Maintenance Agreement
 - What the store should do to boost Maintenance Agreement sales
 - Home Electronics \Rightarrow *
 - What other products should the store stocks up on if the store has a sale on Home Electronics
- Attached mailing in direct marketing
- Detecting "ping-pong"ing of patients
 - transaction: patient
 - item: doctor/clinic visited by a patient
 - support of a rule: number of common patients

4/4/01 collaborative filtering 5

Rule Measures: Support and Confidence

- Find all the rules $X \& Y \Rightarrow Z$ with minimum confidence and support
 - support, s , probability that a transaction contains $\{X, Y, Z\}$
 - confidence, c , conditional probability that a transaction having $\{X, Y\}$ also contains Z .

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Let minimum support 50%, and minimum confidence 50%, we have

- $A \Rightarrow C$ (50%, 66.6%)
- $C \Rightarrow A$ (50%, 100%)

4/4/01 collaborative filtering 6

Mining Association Rules -- Example

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Min. support 50%
Min. confidence 50%

Frequent Itemset	Support
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

For rule $A \Rightarrow C$:

support = support($\{A, C\}$) = 50%

confidence = support($\{A, C\}$) / support($\{A\}$) = 66.6%

The Apriori principle:

Any subset of a frequent itemset must be frequent.

4/4/01

collaborative filtering

7

Mining Frequent Itemsets: the Key Step

- Find the *frequent itemsets*: the sets of items that have minimum support
 - A subset of a frequent itemset must also be a frequent itemset, i.e., if $\{AB\}$ is a frequent itemset, both $\{A\}$ and $\{B\}$ should be a frequent itemset
 - Iteratively find frequent itemsets with cardinality from 1 to k (k -itemset)
- Use the frequent itemsets to generate association rules.

4/4/01

collaborative filtering

8

The Apriori Algorithm

C_k : Candidate itemset of size k

L_k : frequent itemset of size k

$L_1 = \{\text{frequent items}\};$

for ($k = 1; L_k \neq \emptyset; k++$) do begin

C_{k+1} = candidates generated from L_k

for each transaction t in database do
increment the count of all candidates
that are contained in t

L_{k+1} = candidates in C_{k+1} with
min_support

end

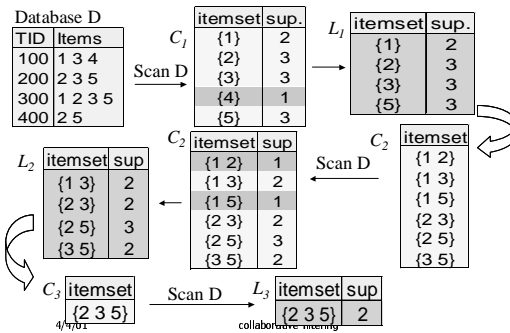
return $\cup_k L_k$

4/4/01

collaborative filtering

9

The Apriori Algorithm -- Example



4/4/01

collaborative filtering

10

Presentation of Association Rules (Table Form)

Rule	Support (%)	Confidence (%)	F	G	H	I
1. cost() = 0.00-1000.00 => revenue() = 0.00-500.00	20.45	40.4				
2. cost() = 0.00-1000.00 => revenue() = 500.00-1000.00	20.45	20.95				
3. cost() = 0.00-1000.00 => order_qty() = 0.00-100.00	59.17	84.94				
4. cost() = 0.00-1000.00 => revenue() = 1000.00-1500.00	10.45	14.84				
5. cost() = 0.00-1000.00 => region() = United States	22.95	32.94				
6. cost() = 1000.00-2000.00 => order_qty() = 0.00-100.00	12.91	69.94				
7. order_qty() = 0.00-100.00 => revenue() = 0.00-500.00	28.45	34.54				
8. order_qty() = 0.00-100.00 => cost() = 1000.00-2000.00	12.91	15.67				
9. order_qty() = 0.00-100.00 => region() = United States	26.9	31.45				
10. order_qty() = 0.00-100.00 => cost() = 0.00-1000.00	59.17	71.95				
11. order_qty() = 0.00-100.00 => product_line() = Tennis	13.92	18.45				
12. order_qty() = 0.00-100.00 => revenue() = 500.00-1000.00	19.67	23.88				
13. product_line() = Tennis => order_qty() = 0.00-100.00	13.92	58.72				
14. region() = United States => order_qty() = 0.00-100.00	25.9	81.54				
15. region() = United States => cost() = 0.00-1000.00	22.95	71.95				
16. revenue() = 0.00-500.00 => cost() = 0.00-1000.00	28.45	100				
17. revenue() = 0.00-500.00 => order_qty() = 0.00-100.00	28.45	100				
18. revenue() = 1000.00-1500.00 => cost() = 0.00-1000.00	10.45	86.75				
19. revenue() = 500.00-1000.00 => cost() = 0.00-1000.00	20.45	100				
20. revenue() = 500.00-1000.00 => order_qty() = 0.00-100.00	19.67	96.14				
21.						
22.						
23. cost() = 0.00-1000.00 => revenue() = 0.00-500.00 AND order_qty() = 0.00-100.00	28.45	40.4				
24. cost() = 0.00-1000.00 => revenue() = 0.00-500.00 AND order_qty() = 0.00-100.00	28.45	40.4				
25. cost() = 0.00-1000.00 => revenue() = 500.00-1000.00 AND order_qty() = 0.00-100.00	19.67	27.93				
26. cost() = 0.00-1000.00 => revenue() = 500.00-1000.00 AND order_qty() = 0.00-100.00	19.67	27.93				
27. cost() = 0.00-1000.00 AND order_qty() = 0.00-100.00 => revenue() = 500.00-1000.00	19.67	33.23				

4/4/01

collaborative filtering

11

Naïve Bayesian Method

- Chapter 4.2 (page 82) of "Data Mining Book" by Witten and Frank
- Instead of computing rules, use training data statistics to get learned model
 - Assumption: attributes are independent of each other given hypothesis
- Then, use the learned model to make prediction

4/4/01

collaborative filtering

12

Example

		Titles				
		Starship Trooper (A)	Sleepless in Seattle (R)	MI-2 (A)	Matrix (A)	Titanic (R)
Users	Sammy	no	yes	no	?	?
	Beatrice	no	yes	no	no	no
	Dylan	no	yes	no	no	yes
	Mathew	yes	no	no	yes	yes
	John	yes	yes	yes	yes	yes

Question: what is the Probability that Sammy likes Matrix?

$$P(\text{Matrix} = \text{yes} | \text{Sammy}) = ?$$

$$P(\text{Matrix} = \text{no} | \text{Sammy}) = ?$$

4/4/01

collaborative filtering

13

Naïve Bayesian Learning

- In general, want to know value hypothesis H (e.g., H=(Matrix=yes))
- Evidence = Attributes = E_i where i goes from 1 to n (number of known purchases for Sammy)
- Bayes rule + Independence assumption

$$P(H | E) = P(E | H) * P(H) / P(E)$$

$$P(E | H) = \prod_i P(E_i | H)$$

- However, $P(E)$ is not known?

4/4/01

collaborative filtering

14

Naïve Bayesian Learning

- Likelihood of "yes"
 - $= P(\text{Sammy} | \text{Matrix} = \text{yes}) * P(\text{Matrix} = \text{yes})$
 - $= P(\text{"Starship"} = \text{no} | \text{Matrix} = \text{yes}) * P(\text{"Seattle"} = \text{yes} | \text{Matrix} = \text{yes}) * P(\text{"MI"} = \text{no} | \text{Matrix} = \text{yes}) * P(\text{Matrix} = \text{yes})$
- $P(\text{"Starship"} = \text{no} | \text{Matrix} = \text{yes}) = 1$
 - To avoid zero probability, use a small number $u=0.01$ to both the numerator and denominator:
 - $P(\text{"Starship"} = \text{no} | \text{Matrix} = \text{yes}) = (0+u/2)/(2+u) = 0.0025$
 - Known as the Laplace Estimator
- $P(\text{"Seattle"} = \text{yes} | \text{Matrix} = \text{yes}) = 0.5$
- $P(\text{"MI"} = \text{no} | \text{Matrix} = \text{yes}) = 0.5$
- $P(\text{Matrix} = \text{yes}) = 0.5$

4/4/01

collaborative filtering

15

Naïve Bayesian Learning (cont)

- Likelihood of "no"
 - $= P(\text{Sammy} | \text{Matrix} = \text{no}) * P(\text{Matrix} = \text{no})$
 - $= P(\text{"Starship"} = \text{no} | \text{Matrix} = \text{no}) * P(\text{"Seattle"} = \text{yes} | \text{Matrix} = \text{no}) * P(\text{"MI"} = \text{no} | \text{Matrix} = \text{no}) * P(\text{Matrix} = \text{no})$
- $P(\text{"Starship"} = \text{no} | \text{Matrix} = \text{no}) = 1$
- $P(\text{"Seattle"} = \text{yes} | \text{Matrix} = \text{no}) = 1$
- $P(\text{"MI"} = \text{no} | \text{Matrix} = \text{no}) = 1$
- $P(\text{Matrix} = \text{no}) = 0.5$

4/4/01

collaborative filtering

16

Finally

- Likelihood of (Matrix=yes) is: $(0.5)^3 = .00125$
- Likelihood of (Matrix=no) is: 0.5
- Thus, do not recommend Matrix to Sammy

4/4/01

collaborative filtering

17

Naïve bayesian based recommendation

- Based on probability
- Can be very accurate when advisors are truly close matches
- Scale up?
- Sparse values \rightarrow too many missing values
- Applications: Charles Ling's work on Direct Marketing (KDD98 paper) (see reference list)

4/4/01

collaborative filtering

18