

Generating Non-Redundant Association Rules

Authors:

Mohammed J. Zaki
Computer Science Department,
Rensselaer Polytechnic Institute,
Troy NY 12180
Zaki@cs.rpi.edu
<http://www.cs.rpi.edu/~zaki>

Overview

- ❖ Introduction And Problem
- ❖ Related Work
- ❖ Key Contribution
- ❖ Methodology
- ❖ Results
- ❖ Opinion

Introduction and Problem

Association Rule Discovery:

- ❖ A successful and important mining task
- ❖ Aims at uncovering all frequent patterns among transaction composed of data
- ❖ Results are presented in the form of rules between sets of items, along with metrics like the joint and conditional probabilities of the antecedent and consequent, to judge rule's importance.

- ❖ Set of Association rules can rapidly grow to be unwieldy, especially as we lower the frequency requirement (Support)
- ❖ Larger the set of frequent item sets more the number of rules
- ❖ Many of them are Redundant.
- ❖ The number of Redundant rules are exponential in the length of longest frequent item set.

Related Work

- ❖ Aclose, Apriori, CMaxMiner

Problem:

- ❖ Some Enumerates frequent item sets.
Produces Redundant rules
- ❖ Some Enumerates Maximal item sets.
Loss of Information
- ❖ Could not be run with lower values of support
- ❖ Generating all subsets of frequent item set takes too much time

Key Contribution

- ❖ CHARM – Algorithm
- ❖ Closed frequent Item Sets

Methodology

Distinct database Items [A, B, C, D]

Transactions

TID	A	B	C	D
TID1	A	B		
TID2	A	B	C	D
TID3	A		C	
TID4	A		C	D
TID5	A	B	C	

All Frequent Itemsets
Minimum support = 40%

Support	Itemsets
100% (5)	A
80% (4)	C, AC
60% (3)	B, AB
40% (2)	D, AD, BC, CD, ABC, ACD

Association Rules
(Confidence = 100%)

Rule	Confiden
BC=>A	(2/2)
D=>AC	(2/2)
AD=>C	(2/2)
CD=>A	(2/2)
D=>C	(2/2)
B=>A	(2/2)
B=>A	(3/3)
C=>A	(4/4)

Frequent Itemset
ACD

Support	Itemsets
100% (5)	A
80% (4)	C, AC
40% (3)	D, AD, CD

Possible Rules: ACD

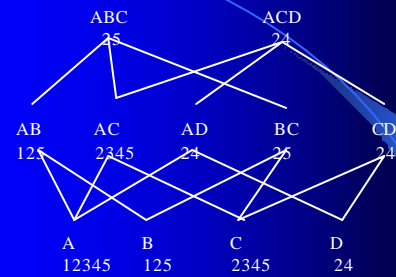
Rule	Confidence
A=>CD	(2/5) 40%
C=>AD	(2/4) 50%
D=>AC	(2/2) 100%
AC=>D	(2/4) 50%
CD=>A	(2/2) 100%
AD=>C	(2/2) 100%
C=>D	(2/4) 50%
D=>C	(2/2) 100%
A=>D	(2/5) 40%
D=>A	(2/2) 100%
A=>C	(4/5) 80%
C=>A	(4/4) 100%

Association Rules

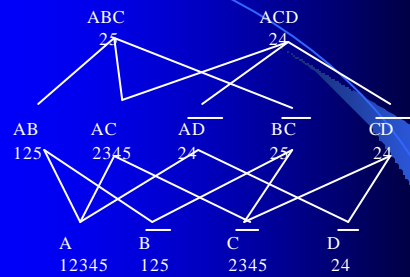
(100% > Confidence <= 80%)

Rule	Confidence
D=>AC	(2/2) 100%
CD=>A	(2/2) 100%
AD=>C	(2/2) 100%
D=>C	(2/2) 100%
D=>A	(2/2) 100%
A=>C	(4/5) 80%
C=>A	(4/4) 100%

Frequent Itemsets



Frequent CLOSED Item sets



Closed Itemset

Definition :

A closed itemset is an itemset X that is the same as its closure.

$X = \text{Cit}(X)$

Similarly,

$Y = \text{Cti}(Y)$ is closed tidset.

The mapping Cit & Cti, being closure operators, satisfy three properties of EXTENSION, MONOTONICITY, IDEMPOTENCY.

Example

Let $X = ABC$

$\text{Cit}(X) = \text{Cit}(ABC)$
 $= i(t(ABC))$
 $= i(25)$
 $= ABC$

Therefore ABC is closed itemset.

Let $X = AD$

$\text{Cit}(X) = \text{Cit}(AD)$
 $= i(t(AD))$
 $= i(24)$
 $= ACD$

Therefore AD is not closed itemset

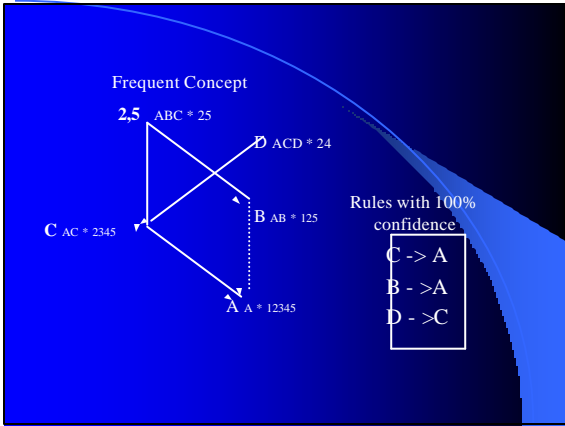
Concept

Closed itemset and Closed tidset pair $X * Y$ is called Concept where

Itemset $X = i(Y)$ and
Tidset $Y = t(X)$

Item Concept
 $C_i(A) = I(t(A)) * t(A)$
 $= I(12345) * 12345$
 $= A * 12345$

Tid Concept
 $C_t(I) = I(I(t(I))) * t(I)$
 $= AB * t(AB)$
 $= AB * 125$



Rule Generation

Redundant rule :
Let R_i denote the rule $X1_i \rightarrow X2_i$. We say that a rule $R1$ is more general than a rule $R2$, denoted $R1 \leq R2$ provided that $R2$ can be generated by adding additional items to either the antecedent or consequent of $R1$.

-Since support of an itemset X equals the support of its closure $C_i(X)$. Thus it is sufficient to consider rules only among the frequent concepts.

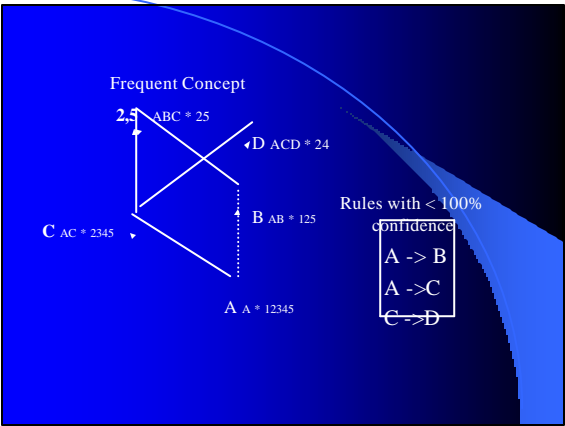
-Observation from concept lattice is that it is sufficient to consider rules among adjacent concepts, since other rules can be inferred by transitivity.

Rules with confidence = 100%

All 100% confidence rules are those that are directed from a super-concept ($X1 * t(X1)$) to a sub-concept ($X2 * t(X2)$), i.e. down arcs.

Example: Item concepts $C_i(C)$ and $C_i(A)$
 $C_i(C) = AC * 2345$
 $C_i(A) = A * 12345$

Therefore $t(C)$ is in $t(A)$
 The rule $C \rightarrow A$ is with 100% confidence.
 Similarly $B \rightarrow A, D \rightarrow C$



General Rules with 100% confidence

$C \rightarrow A$ -----1
 $B \rightarrow A$ -----2
 $D \rightarrow C$ -----3

Other rules with 100% confidence

From 1 and 2, by augmentation
 $BC \rightarrow A$ -----4

From 3 and 1, by transitivity
 $D \rightarrow A$ -----5

From 3 and 5, by augmentation
 $D \rightarrow AC$ -----6

From 1 and 5, by augmentation
 $CD \rightarrow A$ -----7

From 3, adding item A since ACD is closed itemset
 $AD \rightarrow C$ -----8

Rules with Confidence < 100%

These are the rules which are directed from sub-concepts to super-concepts.
 Example: Item concepts $C_i(A)$ and $C_i(B)$
 $C_i(A) = A * 12345$
 $C_i(B) = AB * 125$
 Therefore $t(B)$ is in $t(A)$
 The rule $A \rightarrow B$ is with 60% confidence.
 Similarly $A \rightarrow C$ is with 80% confidence,
 $C \rightarrow D$ is with 50% confidence.

Complexity of rule generation

Traditional Framework : $O(f \cdot 2^l)$
 where f = total number of frequent itemsets
 l = length of longest frequent itemsets
 In Example ,
 $f = 11, l = 3$
 Number of rules generated = $O(f \cdot 2^l) = 88$

With new Framework
 $f = 5, l = 3$
 Number of rules generated = $O(f \cdot 2^l) = 40$

Results

Database	Sup	Length	No of frequent item sets	No. of closed frequent item sets	ratio	Time taken Apriori	Time taken CHARM	Ratio
chess	80%	10	8227	5083	1.6	18.54	1.92	9.7
chess	70%	13	48969	23991	2	213.03	8.17	26.1
connect	97%	6	487	284	1.7	19.7	4.15	4.7
connect	90%	12	27127	3486	7.8	2084.3	43.8	47.6
mushroom	40%	7	565	140	4	1.56	0.28	5.6
mushroom	20%	15	53583	11974	4.7	167.5	1.2	144.4
pumsb*	60%	7	167	68	2.5	11.4	1	11.1
pumsb*	40%	13	27354	2610	10.5	847.9	174	49.6
pumsb	95%	5	172	110	1.6	19.7	1.7	11.7
pumsb	85%	10	20533	8513	2.4	1379.8	76.1	18.1
T20112D100K	0.5%	9	2890	2067	1.4	6.3	5.1	1.2
T4018D100K	1.5%	13	12088	4218	2.9	41.6	15.8	2.6
T1014D100K	0.5%	5	1073	1073	1	2	1.1	1.8
T1014D100K	0.1%	10	27532	26806	1.03	32.9	8.3	4
T2014D100K	1.0%	6	1327	1327	1	6.7	4.8	2.6

Number of Item sets and running time

Database	Sup	Length	All possible rules Traditional	Closed	Ratio	Rules with one consequent	Ratio
chess	80%	10	882564	27711	20	44631	2
chess	70%	13	8171198	152074	5.4	318248	2
connect	97%	6	8092	3116	7	1846	1.7
connect	90%	12	3640704	18548	19.3	170067	9
mushroom	40%	7	7020	475	1.5	1906	4
mushroom	20%	15	19191656	5741	3343	389999	66
pumsb*	60%	7	2338	192	12	526	3
pumsb*	40%	13	5659536	13479	420	179638	13
pumsb	95%	5	1170	267	4	473	2
pumsb	85%	10	1408950	44483	32	113089	3
T20112D100K	0.5%	9	40356	2642	1.5	6681	3
T4018D100K	1.5%	13	1609678	11379	142	63622	6
T1014D100K	0.5%	5	2216	1231	1	1231	1
T1014D100K	0.1%	10	431838	86902	5	90350	1.04
T2014D100K	1.0%	6	2736	1738	1	1738	1
T2014D100K	0.25%	10	391512	89963	4	90911	1.01

Number of Rules (all vs. consequent of length 1) (Sup=minsup, Len=longest itemset)

Opinion

This paper mainly focus on reducing redundant rules and it imposes on more general rules instead of finding all possible rules. This paper uses concept of closed frequent item set, instead of frequent item sets.

Result shows that number of closed frequent item set are less compared to frequent itemsets and so the less number of rules are generated. And at the same time we don't loose any information.

So Closed frequent item sets are very efficient in finding non-redundant rules.

We rank this paper 7 on scale of 0 to 10, 10 as best.