Efficient Mining of Dissociation Rules

Mikołaj Morzy

7^{th} International Conference DaWaK 2006
Kraków, Poland, September 2006

Outline

Mining "negative knowledge"

- association rules capture only "positive knowledge"
  $'wine' \wedge 'grapes' \Rightarrow 'cheese' \wedge 'white\ bread'$
- what about "negative knowledge"?
  $'FC\ Barcelona\ jersey' \Rightarrow \neg\, 'Real\ M.\ scarf' \wedge \neg\, 'Real\ M.\ cup'$
- ...or another type of "negative pattern"?
  $'beer' \wedge 'sausage' \Rightarrow 'mustard' \wedge \neg\, 'red\ wine'$

Mining “negative knowledge”

- association rules capture only “positive knowledge”
  $'wine' \wedge 'grapes' \Rightarrow 'cheese' \wedge 'white\ bread'$
- what about “negative knowledge”?
  $'FC\ Barcelona\ jersey' \Rightarrow \neg 'Real\ M.\ scarf' \wedge \neg 'Real\ M.\ cup'$
- …or another type of “negative pattern”?
  $'beer' \wedge 'sausage' \Rightarrow 'mustard' \wedge \neg 'red\ wine'$

Observation

Mining of “negative knowledge” is difficult due to

- sparsity of data
- unmanageable number of association rules with negation

Where is the problem?

Recall the definition of data mining
"...discovery and extraction of non-trivial, ultimately understandable, previously unknown, valid, useful and utilitarian patterns from large data volumes" (Shapiro et al.)

Where is the problem?

Recall the definition of data mining
"...discovery and extraction of non-trivial, ultimately understandable, previously unknown, valid, useful and utilitarian patterns from large data volumes" (Shapiro et al.)

Observation

What is wrong with current solutions?

- too complex
- models are too big
- not useful in practice

Illustration of the problem

id	items
1	A B D
2	B C
3	A D E
4	B D E
5	A B C

Illustration of the problem

id	items
1	A B D
2	B C
3	A D E
4	B D E
5	A B C

$minsup = 40\%$, there are 9 frequent itemsets

$$L_D = \{A, B, C, \ldots, BC, BD\}$$

Illustration of the problem

id	items
1	A B D
2	B C
3	A D E
4	B D E
5	A B C

$minsup = 40\%$, there are 9 frequent itemsets

$$L_D = \{A, B, C, \ldots, BC, BD\}$$

$minsup = 40\%$, there are 34 (!) frequent itemsets with negation

$$L'_D = \{A, A', B, C, C', \ldots, AB, AC', AD, \ldots, BCD'E'\}$$

Our solution

Enter the dissociation rules

- find negatively associated sets of items while keeping the number of discovered patterns low
- simplicity over sophistication
- sacrifice the abundance of patterns for actionability and usefulness of the result

Our solution

Enter the dissociation rules

- find negatively associated sets of items while keeping the number of discovered patterns low
- simplicity over sophistication
- sacrifice the abundance of patterns for actionability and usefulness of the result

Contribution

- introduction of dissociation rules formalism
- development of the DI-Apriori algorithm
- experimental evaluation of the proposal

Related Work

- association rules (Agrawal et al.): $A \wedge B \Rightarrow C$
- excluding associations (Amir et al.): $A \wedge \neg B \Rightarrow C$
- unexpected association rules (Savasere et al.): taxonomy, expected support
- confined negative association rules (Antonie et al.): $A \Rightarrow \neg B$, $\neg A \Rightarrow B$, $\neg A \Rightarrow \neg B$
- generalized negative association rules (Kryszkiewicz et al.): derivable and non-derivable itemsets, certain rules, negative border, rule generators
- unexpected patterns (Padmanabhan et al.): background knowledge, expectations and beliefs
- exception rules (Liu et al.): unexpected deviation from a well-established fact

Basic Definitions

- set of items $I = \{i_1, \ldots, i_n\}$, database D, $\forall t_i \in D : t_i \subseteq I$
- transaction t *supports* an item x if $x \in t$
- transaction t *supports* an itemset X if $\forall x \in X : x \in t$
- *support* of an itemset X, denoted $support_D(X)$, is the number of transactions in D supporting the itemset
- itemset X is a *frequent itemset* if $support_D(X) \geq minsup$
- given $X, Y \subset I$, *support* of an itemset $\{X \cup Y\}$ is called the *join* of X and Y

Basic Definitions

- given a collection L_D of frequent itemsets in D, the *negative border* $Bd^-(L_D)$ of the collection of frequent itemsets consists of minimal itemsets not contained in L_D, $Bd^-(L_D) = \{X : X \notin L_D \wedge \forall Y \subset X, Y \in L_D\}$
- given user-defined thresholds *minsup* and *maxjoin*, where $minsup > maxjoin$
- itemset Z is a *dissociation itemset* if $support_D(Z) \leq maxjoin$ and itemsets X, Y exist, such that $support_D(X) \geq minsup$, $support_D(Y) \geq minsup$, and $X \cup Y = Z$

Basic Definitions

Dissociation Rule

An expression $X \not\Rightarrow Y$, where $X \subset I$, $Y \subset I$, $X \cap Y = \emptyset$

- $support_D(X \cup Y) \leq maxjoin$
- $support_D(X) \geq minsup$
- $support_D(Y) \geq minsup$
- X is the *antecedent* of the rule
- Y is the *consequent* of the rule
- $X \not\Rightarrow Y$ is a *minimal dissociation rule* if $\nexists X' \subseteq X, Y' \subseteq Y$ such that $X' \not\Rightarrow Y'$ is a valid dissociation rule

Basic Measures

$$support_D(X \not\Rightarrow Y) = \min\{support_D(X), support_D(Y)\}$$

Basic Measures

$$support_D(X \not\Rightarrow Y) = \min\{support_D(X), support_D(Y)\}$$

$$join_D(X \not\Rightarrow Y) = support_D(X \cup Y)$$

Basic Measures

$$support_D(X \not\Rightarrow Y) = \min\{support_D(X), support_D(Y)\}$$

$$join_D(X \not\Rightarrow Y) = support_D(X \cup Y)$$

$$\begin{aligned} confidence_D(X \not\Rightarrow Y) &= \frac{support_D(X) - support_D(X \cup Y)}{support_D(X)} = \\ &= 1 - \frac{join_D(X \not\Rightarrow Y)}{support_D(X)} \end{aligned}$$

Problem Formulation

Given a database D and thresholds of minimum support, confidence, and maximum join, called *minsup*, *minconf*, and *maxjoin*, respectively. Find all dissociation rules valid in the database D with respect to the above mentioned thresholds

Thresholds

User-defined thresholds are used as follows:

- *minsup* selects statistically significant itemsets for antecedents and consequents of generated dissociation rules
- *maxjoin* provides an upper limit of antecedent and consequent co-occurrence in the database
- *minconf* post-processes discovered dissociation rules in search for strong dissociations

note the lower bound $\underline{confidence_D} = (1 - maxjoin/minsup)$

Lemmas

Lemma 1. Let L_D denote the set of frequent itemsets discovered in the database D. If $X \not\Rightarrow Y$ is a valid dissociation rule, then $(X \cup Y) \notin L_D$

Lemmas

Lemma 1. Let L_D denote the set of frequent itemsets discovered in the database D. If $X \nRightarrow Y$ is a valid dissociation rule, then $(X \cup Y) \notin L_D$

Lemma 2. If $X \nRightarrow Y$ is a valid dissociation rule, then $\forall X' \supseteq X, Y' \supseteq Y$ such, that $X' \in L_D \wedge Y' \in L_D$, $X' \nRightarrow Y'$ is a valid dissociation rule

Lemmas

Lemma 1. Let L_D denote the set of frequent itemsets discovered in the database D. If $X \not\Rightarrow Y$ is a valid dissociation rule, then $(X \cup Y) \notin L_D$

Lemma 2. If $X \not\Rightarrow Y$ is a valid dissociation rule, then $\forall X' \supseteq X, Y' \supseteq Y$ such, that $X' \in L_D \wedge Y' \in L_D$, $X' \not\Rightarrow Y'$ is a valid dissociation rule

Lemma 3. $\forall X, Y$ such, that $X \not\Rightarrow Y$ is a valid dissociation rule, there exists $Z \in Bd^{-}(L_D)$ such, that $(X \cup Y) \supseteq Z$

Naive Approach

1. find the collection L_D of frequent itemsets using Apriori algorithm
2. join all possible pairs of frequent itemsets to form candidate dissociation itemsets
3. prune candidate dissociation itemsets contained in L_D based on Lemma 1.
4. count the support of candidate dissociation itemsets during a full database scan
5. generate dissociation rules

DI-Apriori

From Lemma 2 follows that it is sufficient to discover only minimal dissociation rules
From Lemma 3 follows that the search space is limited to supersets of sets from the negative border $Bd^{-}(L_D)$

Notation

- L_D^1: the set of frequent 1-itemsets
- $C_{\not\Rightarrow}$: the set of pairs of frequent itemsets that are candidates for joining into a dissociation itemset
- $D_{\not\Rightarrow}$: the set of pairs of frequent itemsets that form valid dissociation itemsets

DI-Apriori

1. form initial candidate dissociation itemsets ($C_{\not\Rightarrow}$) based on the negative border $Bd^{-}(L_D)$
2. extend candidate dissociation itemsets with frequent 1-itemsets from L_D^1
3. compute the support of candidate dissociation itemsets and prune them on *maxjoin*
4. extend dissociation itemsets ($D_{\not\Rightarrow}$) with frequent supersets of their antecedents and consequents
5. compute the support of dissociation itemsets, if necessary
6. generate dissociation rules

Comparison of Algorithms

- Naive approach: single database scan, many candidate dissociation itemsets
- DI-Apriori: few database scans, few candidate dissociation itemsets

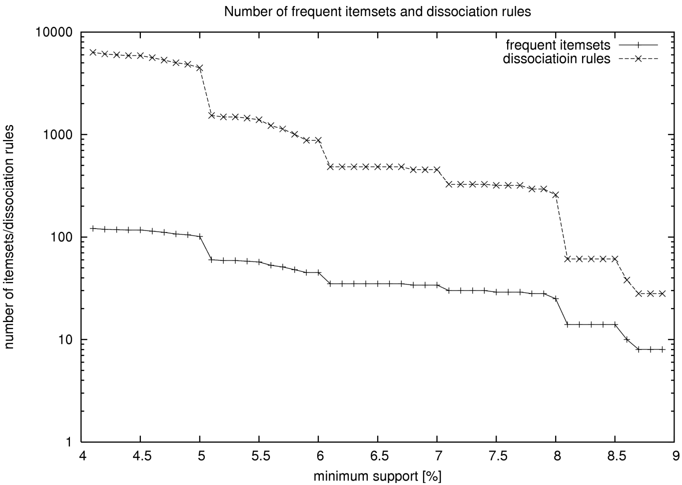
Table: Number of itemsets processed by Basic Apriori vs. DI-Apriori

minsup	*maxjoin*	Basic Apriori frequent	Basic Apriori candidate	DI-Apriori
5%	1%	83	396	264
4%	1%	214	2496	1494
3%	1%	655	16848	4971

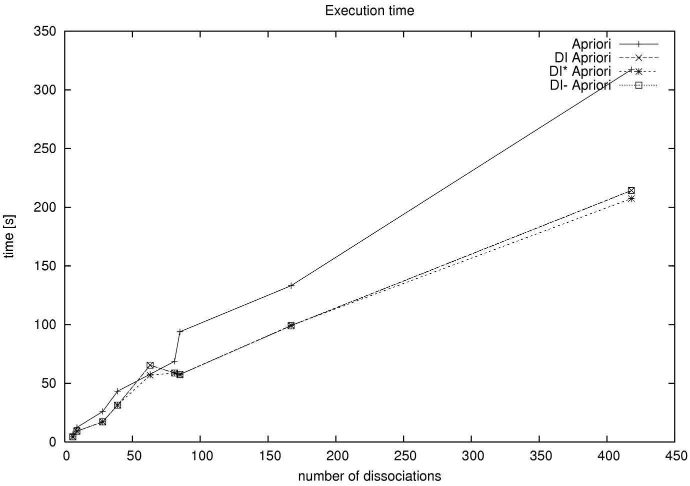
Synthetic Datasets

- DBGen generator from IBM's Quest Project
- number of transactions: 20 000
- average transaction size: 10 items
- number of patterns: 300
- average pattern size: 4 items
- *maxjoin* threshold: 3% (if not stated otherwise)
- *minsup* threshold: 5% (if not stated otherwise)

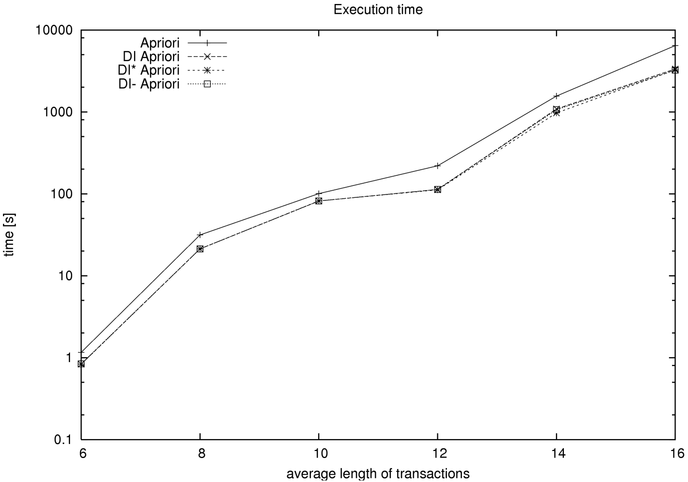
Number of frequent itemsets and dissociation rules



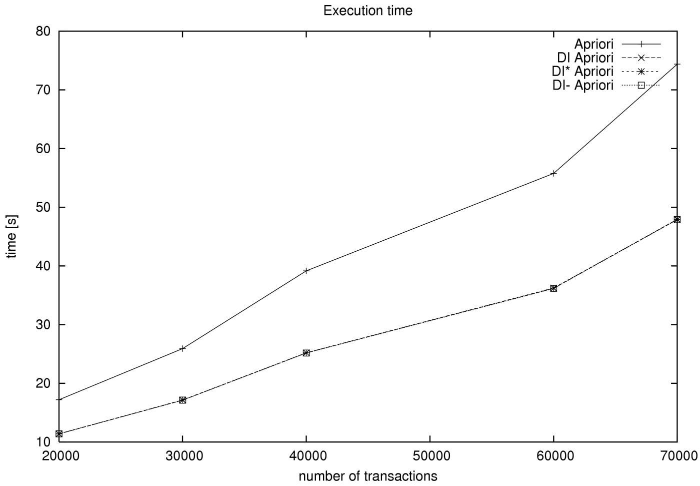
Execution time w.r.t the number of dissociation rules



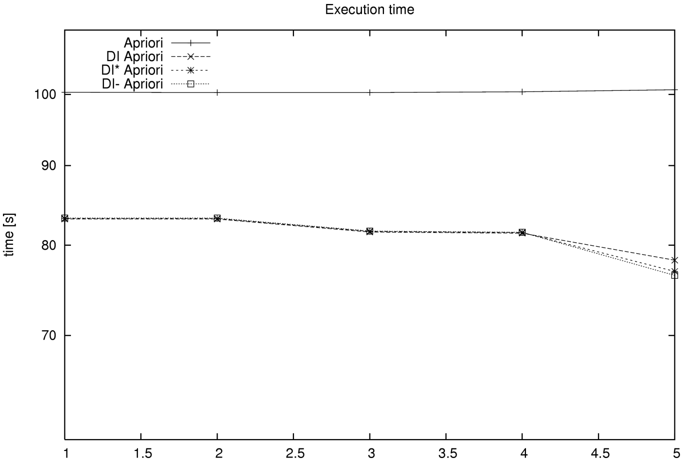
Execution time w.r.t. the average length of transaction



Execution time w.r.t. the number of transactions



Execution time w.r.t. the gap between *minsup* and *maxjoin*



Conclusions and Future Work

Conclusions

- initial research on dissociation rules
- simple model that captures "negative" knowledge
- main advantages: simplicity, practical feasibility, usability

Conclusions and Future Work

Future Work

- experimental comparison with other types of “negative” association rules
- behavior on real-world data sets
- development of concise and compact representations of dissociation rules