Parallelizing the Mining of Frequent Itemsets in Uncertain Data

Erich A. Peterson and Peiyi Tang Department of Computer Science The University of Arkansas at Little Rock

Overview

Within traditional (certain) databases, an item either occurs or it does not. Given a set of items $I = \{x_1, x_2, \ldots, x_m\}$, an itemset $X \subseteq I$, and a set of transactions $T = \{t_1, t_2, \ldots, t_n\}$—the support of an itemset (denoted $Sup_T X$) is the # of transactions that contain X. Ex. $Sup_T(b, c) = 2$.

	a	b	c
t_0	x		x
t_1	x	x	x
t_2		x	
t_3		x	x

	a	b	c
t_0	0.9		0.21
t_1	0.45	1.0	0.34
t_2		0.88	
t_3		0.6	0.4

However, with uncertain databases, each item x has a probability of being in transaction t_j denoted as $P(x \in t_j)$. Ex. $P(a \in t_1) = 0.45$.

Uncertain Data Model

If items and transactions are independent the following gives the prob. of a possible world w:

$$P(w) = \prod_{t \in T(w)} \left(\prod_{x \in t} P(x \in t') \cdot \prod_{x \notin t} (1 - P(x \in t')) \right)$$

where $T(w)$ is the set of certain transactions of world w, t a certain transaction in $T(w)$, t' the corresponding *uncertain* transaction in uncertain database T, and $P(x \in t')$ the existential probability of item x in the uncertain transaction t'.

Thus, we could calc. the probability of itemset X having support i as follows:

$$P_i(X) = \sum_{w \in W, Sup_{T(w)}(X) = i} P(w)$$

where W is the set of possible worlds.

Bernecker et al. showed you can calc. it as:

$$P_i(X) = \sum_{S \subseteq T, |S| = i} \left(\prod_{t \in S} P(X \subseteq t) \cdot \prod_{t \in T - S} (1 - P(X \subseteq t)) \right)$$

where T is the original uncertain database and $P(X \subseteq t)$ is

$$P(X \subseteq t) = \prod_{x \in X} P(x \in t)$$

Thus, the probability of the support of X being at least i is:

$$P_{\geq i}(X) = \sum_{k=i}^{|T|} P_k(X)$$

Thus, $P_{\geq minsup}(X)$ is the probability that X is frequent, and if this value is above a user-defined confidence threshold τ, then X is considered a *probabilistic frequent itemset*. Ex. $P_{\geq minsup}(X) \geq \tau$

Probabilistic Frequent Calculation

Li et al. (in the context of probabilistic ranking) presented the concept of generation functions, which Bernecker et al. applied to calculating $P_{\geq minsup}(X)$. Given the generating function:

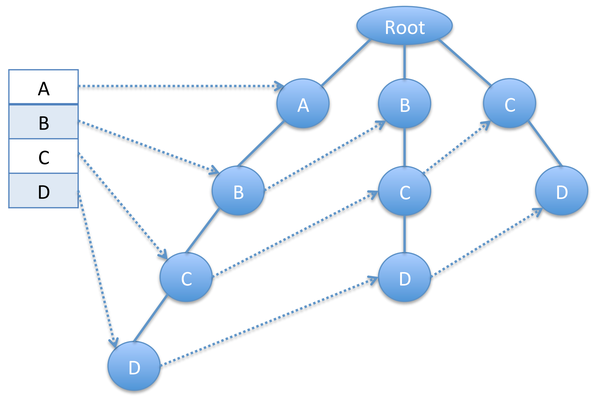
$$F^i = \prod_{t=1}^{i} (1 - P(X \subset t) + P(X \subset t) \cdot x) = \sum_{j=0}^{i} c_j$$

where each coefficient c_j in F^i is the probability that the support of X is j in the first i transactions. Thus,

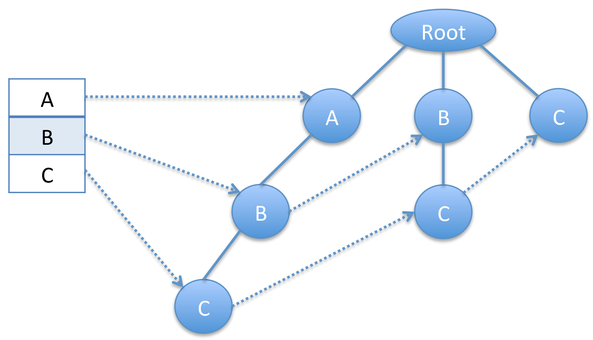
$$P_{\geq minsup}(X) = \sum_{j=minsup}^{|T|} c_j, \text{where each } c_j \text{ is from } F^{minsup}.$$

Mining and Data Structure

Itemsets that share a common prefix can share nodes in this adaptation of the FP-Tree. Information about probabilities of each item is encoded in each node. After an item is determined to be probabilistically frequent a projection or conditional tree is recursively created dependent upon the existence of that item. That is, statistics on that item are propagated upward.



Ex: If "D" is probabilistically frequent, then the tree below is created based on the fact that. In addition, "D"s statistics are propagated upward, and nodes along that path retain only those statistics in which "D" and the associated node co-exist. If then that tree is mined and "C" is found to be frequent, then we deduce that the itemset {D, C} is probabilistically frequent. Notice, that the search space can be partitioned into those itemsets that begin with each singleton item.



Ex: If 4 items exist (A, B, C, D), then 4 jobs can be created, each only concerned with mining itemsets which start with their given singleton.

Conclusions & Future Work

We have explored a way to parallelize the mining algorithm used by Bernecker et al. to discover probabilistic frequent itemsets in uncertain data. The program is currently under development and will utilize Condor-G for the distribution of work.

Next, we plan on extending this algorithm to mine probabilistic itemsets in an incremental fashion. That is, given a database that is evolving over time (adding and deleting transactions).