THE 18TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING
KDD 2012 BEIJING
Beijing, China August 12-16, 2012

UFIMT: An Uncertain Frequent Itemset Mining Toolbox

Yongxin Tong
HKUST
yxtong@cse.ust.hk

Lei Chen
HKUST
leichen@cse.ust.hk

Philip S. Yu
University of Illinois at Chicago
psyu@cs.uic.edu

Motivations

Unlike the frequent itemset in deterministic data has a unique definition, the frequent itemset under uncertain environments has two different definitions. Most existing works focus on one of the definitions and no comprehensive study is conducted to compare the two different definitions. Thus, we hope to

1. Present a novel prototype system, *UFIMT*, for mining frequent itemsets over uncertain databases.
2. Clarify the relationship of the two definitions of frequent itemsets over uncertain data.
3. Provide uniform baseline implementations for the existing representative algorithms for mining frequent itemsets under uncertain databases.
4. Propose an objective and sufficient experimental evaluation platform.

Basic Concepts

Definition 1 (Expected Support). Given an uncertain database *UDB* which includes *N* transactions, and an itemset *X*, the expected support of *X* is:

$$esup(X) = \sum_{i=1}^{N} p_i(X)$$

where $p_i(X)$ is the probability of the itemset *X* appearing in the i-th transaction.

Definition 2 (Expected-Support-based Frequent Itemset). Given an uncertain database *UDB*, and a minimum expected support ratio, *min_esup*, an itemset *X* is an expected support-based frequent itemset if and only if $esup(X) > N \times min\_esup$.

Definition 3 (Frequent Probability). Given an uncertain database *UDB* which includes *N* transactions, a minimum support ratio *min_sup*, and an itemset *X*, *X's* frequent probability, denoted as Pr(*X*), is:

$$\Pr(X) = \Pr\{\sup(X) \geq N \times min\_sup\}$$

Definition 4 (Probabilistic Frequent Itemset). Given an uncertain database *UDB*, a minimum support ratio *min_sup*, and a probabilistic frequent threshold *pft*, an itemset X is a probabilistic frequent itemset if X's frequent probability is larger than the probabilistic frequent threshold, namely, $\Pr(X) > pft$.

Type	Algorithms	Highlights
Expected Support–based Frequent Algorithms	UApiori	Apriori-based search strategy
	UFP-growth	UFP-tree index structure ; Pattern growth search strategy
	UH-Mine	UH-struct index structure ; Pattern growth search strategy
Exact Probabilistic Frequent Algorithms	DP	Dynamic programming-based exact algorithm
	DC	Divide-and-conquer-based exact algorithm
Approximation Probabilistic Frequent Algorithms	PDUApiori	Poisson-distribution-based approximation algorithm
	NDUApiori	Normal-distribution-based approximation algorithm
	NDUH-Mine	Normal-distribution-based approximation algorithm; UH-struct index structure

Table 1. A Summary of Representative Algorithms in UFIMT

Architecture and Demonstration

The system architecture and a screenshot of UFIMT is shown in Figure 1 and Figure 2. Moreover, two figures of experimental comparisons over UFIMT are illustrated in Figure 3 and 4.

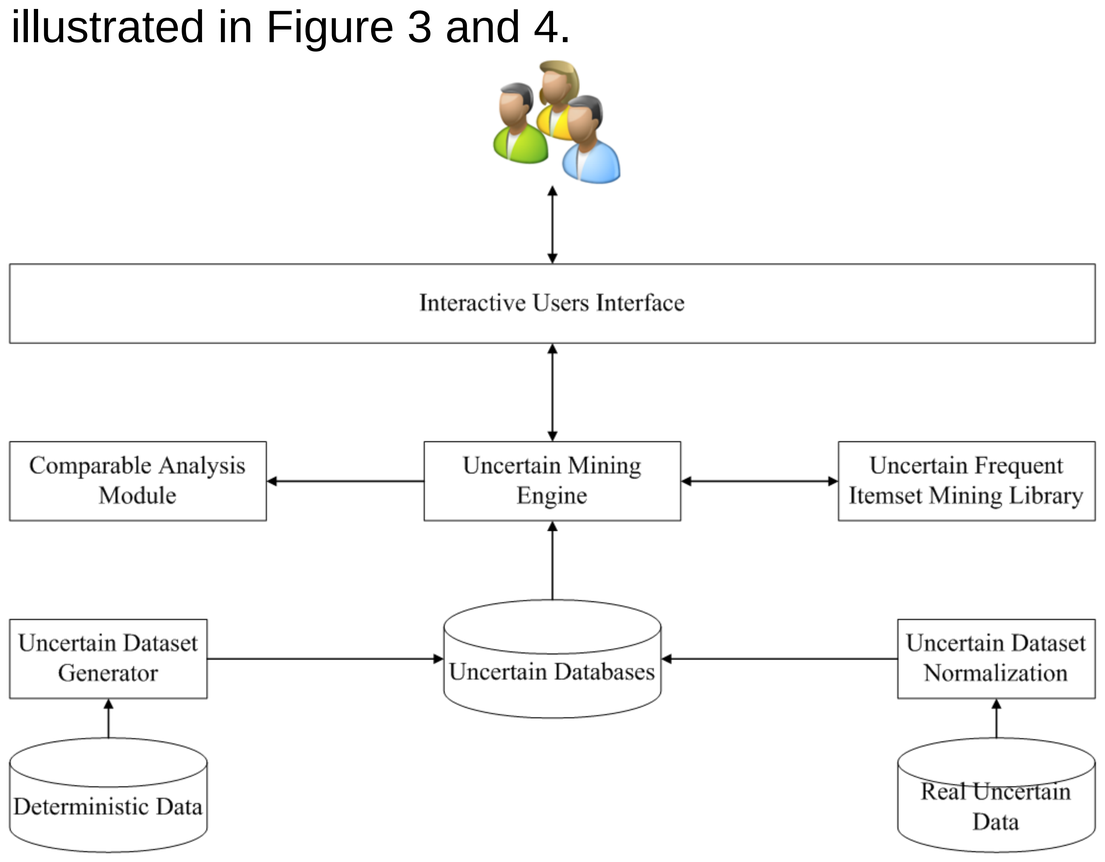

Figure 1. The System Architecture of UFIMT

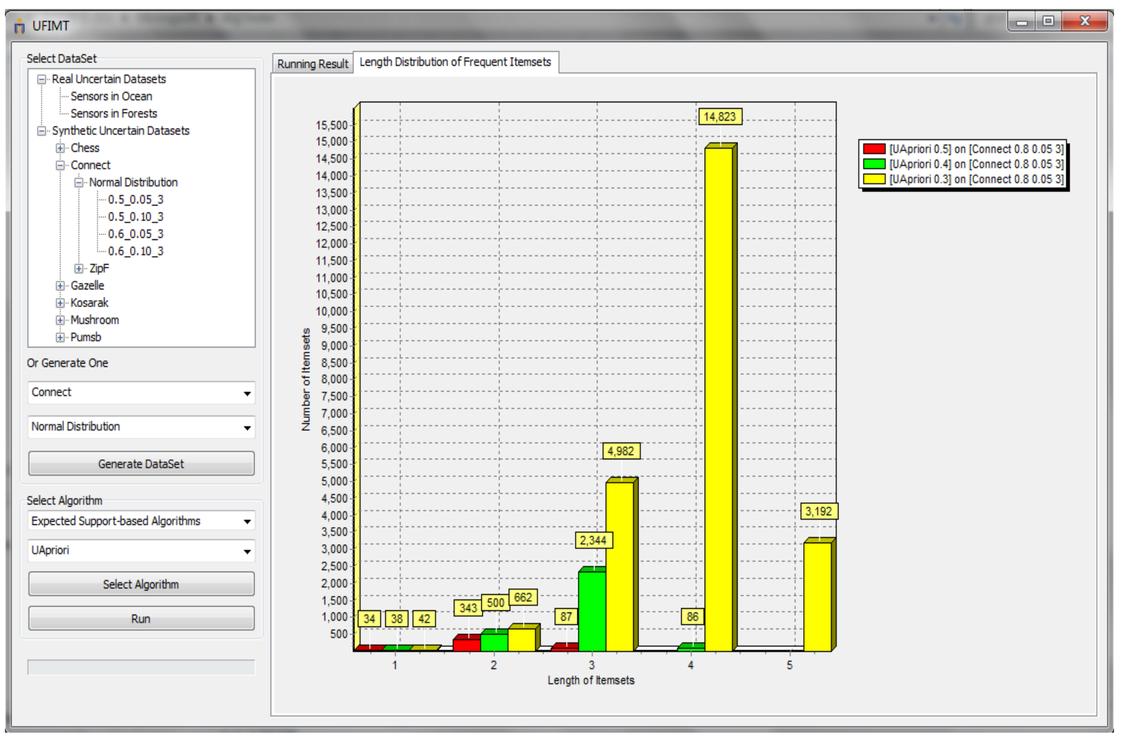

Figure 2. Length distribution of frequent itemsets vs. *min_esup*

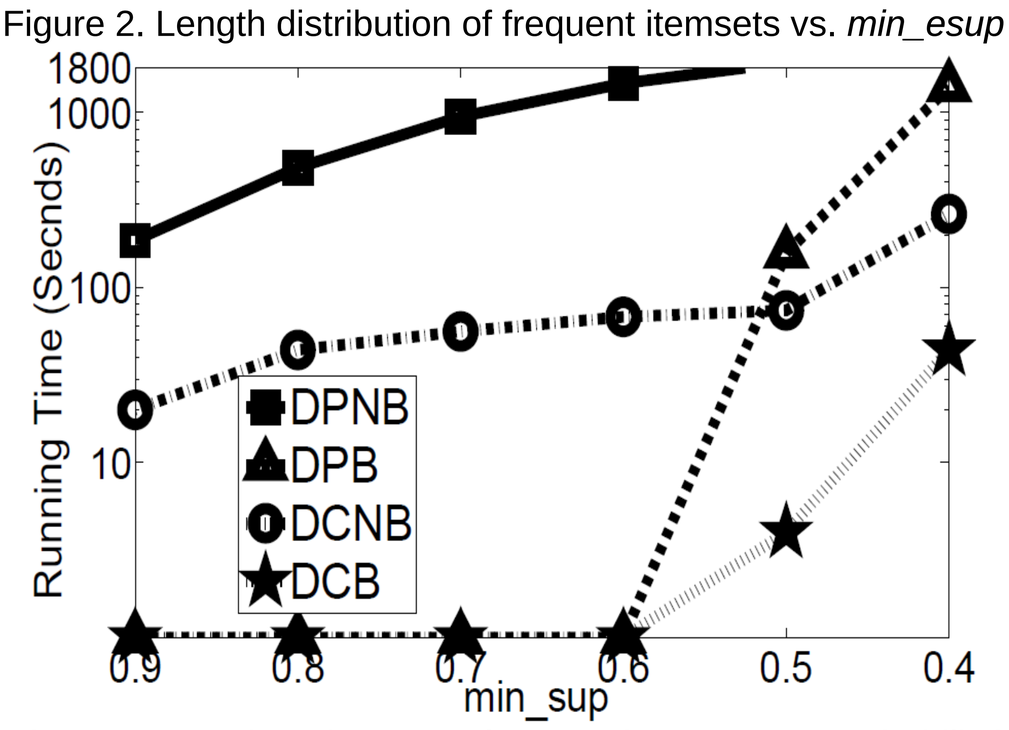

Figure 3. Running Time vs. *min_sup* of Four Exact Probabilistic Frequent Algorithms in *Accident* Dataset

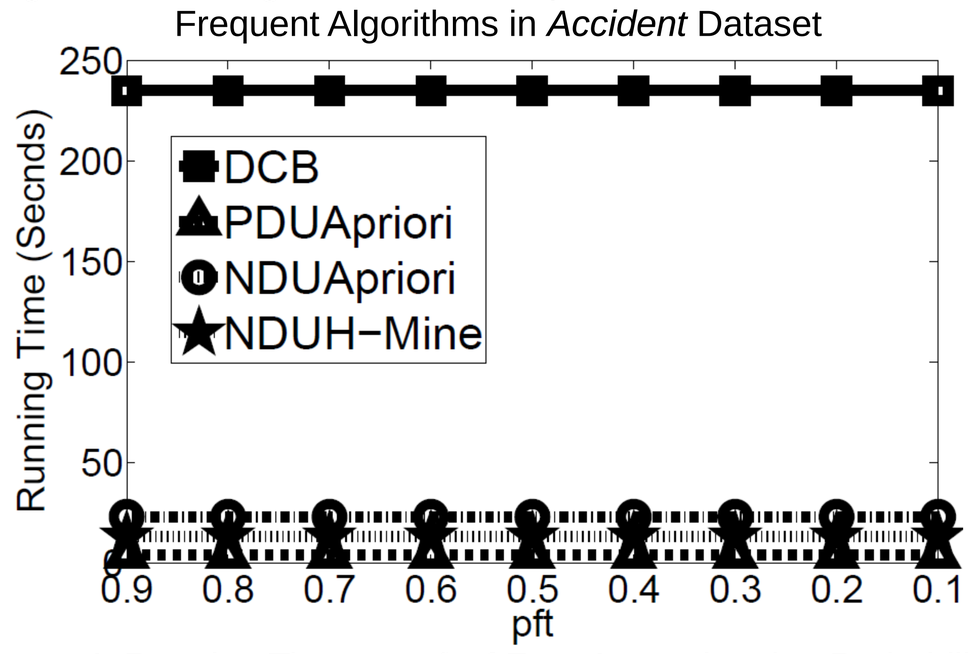

Figure 4. Running Time vs. *pft* of Four Approximation Probabilistic Frequent Algorithms in *Kosarak* Dataset

Paper ID: de1210