GPU-based acceleration of data mining algorithms

Toshiyuki Amagasa
Center for Computational Sciences
University of Tsukuba

Talk outline

- Database group at CCS
- Overview of research topics
- GPU-based acceleration of frequent itemset mining from uncertain databases
- GPU-based acceleration of canopy clustering
- GPU-based acceleration of uncertain time series search
- Future collaboration

Database group at CCS

CCS

Hiroyuki Kitagawa
Professor
CCS & Dept. CS
- Database
- Data mining

Toshiyuki Amagasa
Assoc. Prof.
CCS & Dept. CS
- Database
- Data mining

Dept. CS

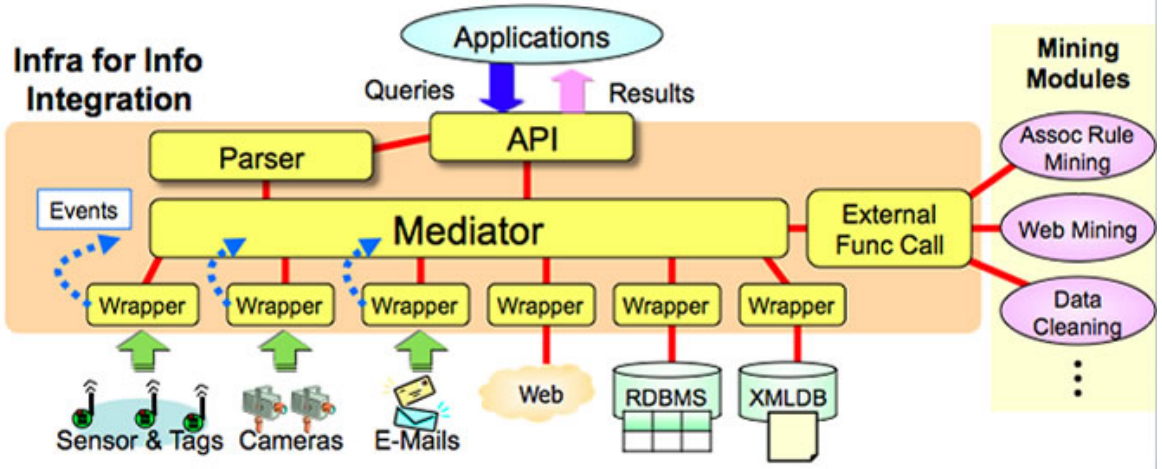
Yasuhiro Hayase
Assistant Prof.
Dept. CS
- Software engineering
- Repository mining

Chiemi Watanabe
Assistant Prof.
Dept. CS
- Database
- Data privacy

Postdoc x2, D x5, M x17 (+6), B x7
Research student x4

Research topics 1

- Infrastructure for information integration
  - Data stream processing
  - Integration of data streams and heterogeneous information sources



- Data mining / social media mining
  - Outlier detection
  - Mining from Twitter

Research topics 2

- GPU-based acceleration of data mining
  - Frequent itemset mining
  - Time series search
  - Clustering
- Web information systems
  - XML databases
  - RDF/LOD databases
- Database applications in scientific domains
  - GPV/JMA archive
  - JLDG/ILDG
  - Biological database

GPU-based Frequent Itemset Mining over Uncertain Databases

Yusuke Kozawa, Toshiyuki Amagasa,
Hiroyuki Kitagawa

University of Tsukuba

Uncertain Transaction Databases

- Transaction databases
  - Purchase records
  - Observation records
  - System logs
- Uncertainty
  - Each transaction has an existential probability.
  - The probability specifies the chance that the transaction exists.

ID	Itemset	Prob.
T1	{game, music}	0.5
T2	{music, video}	0.7
T3	{game}	0.8
T4	{music}	0.9

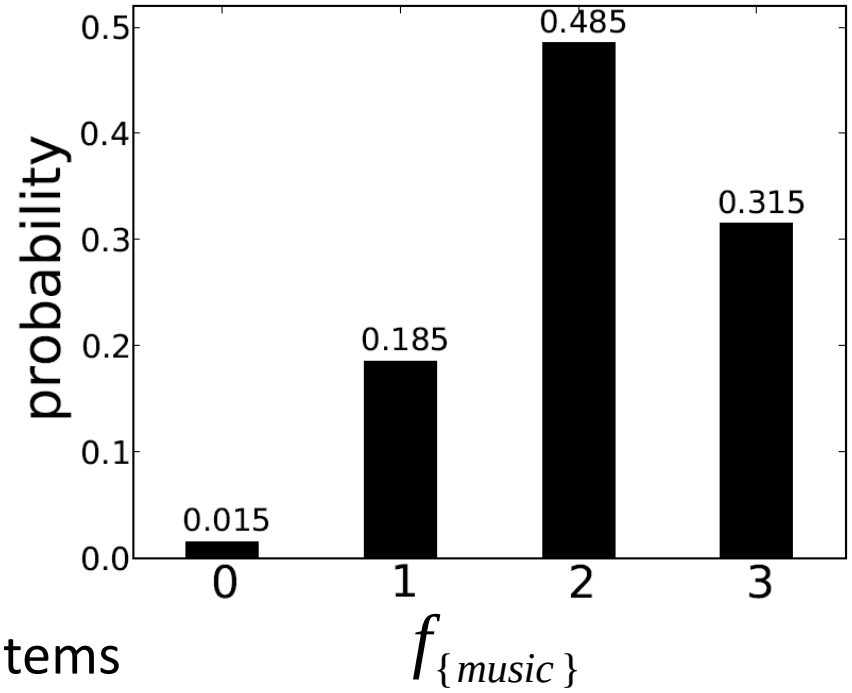
Frequent itemset mining

- Frequent itemset mining is to find frequently occurring patterns from a transaction database.
  - Find characteristic patterns from
    - Purchase / observation records
    - System logs
- To find frequent itemsets from uncertain databases, we need to care about the uncertainty.

Support Probability Mass Function

- Support Probability Mass Function (SPMF) f_X
  - The probability mass function of sup(X)

ID	Itemset	Prob.
T1	{game, music}	0.5
T2	{music, video}	0.7
T3	{game}	0.8
T4	{music}	0.9



➔ More complicated when many items

Probabilistic Frequent Itemsets

- Probabilistic Frequent Itemset (PFI) *X*

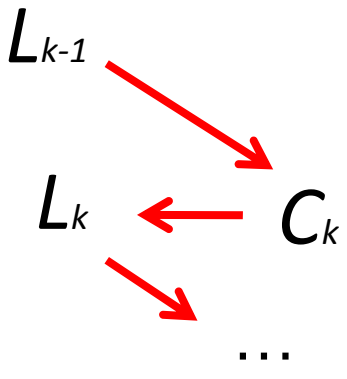
$$P(\sup(X) \geq minsup) \geq minprob$$

The probability that *X* is a frequent itemset $= \sum_{k=minsup}^{n} f_X(k)$

  - Minsup and minprob are the support threshold and the probability threshold respectively
- The problem: probabilistic frequent itemset mining
  - Given an uncertain transaction databases, minsup, and minprob, return all PFIs

pApriori Algorithm [Sun et al., KDD '10]

- Inputs: uncertain transaction database, minsup, and minprob
- This algorithm consists of two procedures
  1. Generate size-k Candidate PFIs C_k from size-(k-1) PFIs L_{k-1}
  2. Extract size-k PFIs from the size-k Candidate PFIs C_k

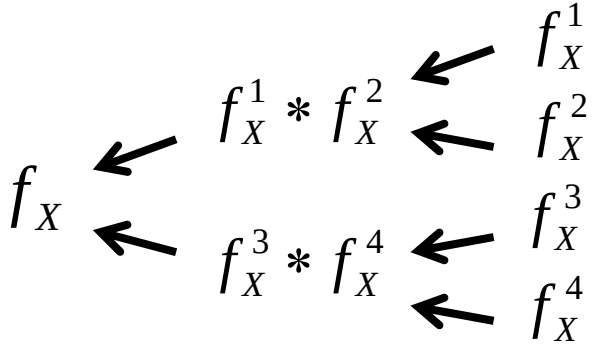


This needs to compute the SPMFs

Since this is the most computationally expensive step, it's important to efficiently compute the SPMFs

Efficient Computation of SPMFs using a GPU

- Convolution



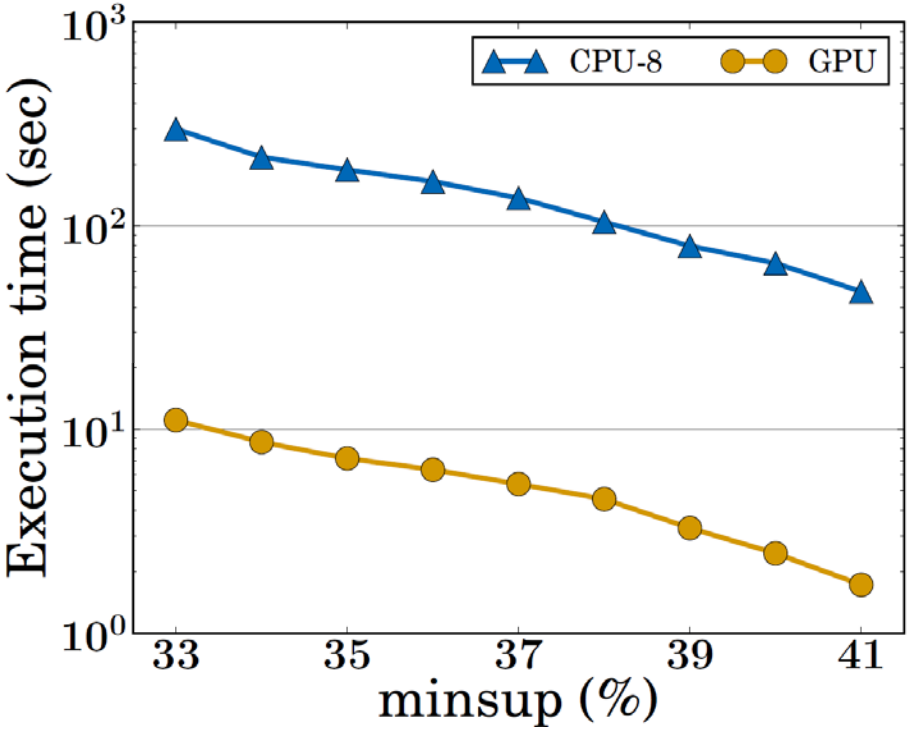
ID	Itemset	Prob.
T1	{game, music}	0.5
T2	{music, video}	0.7
T3	{game}	0.8
T4	{music}	0.9

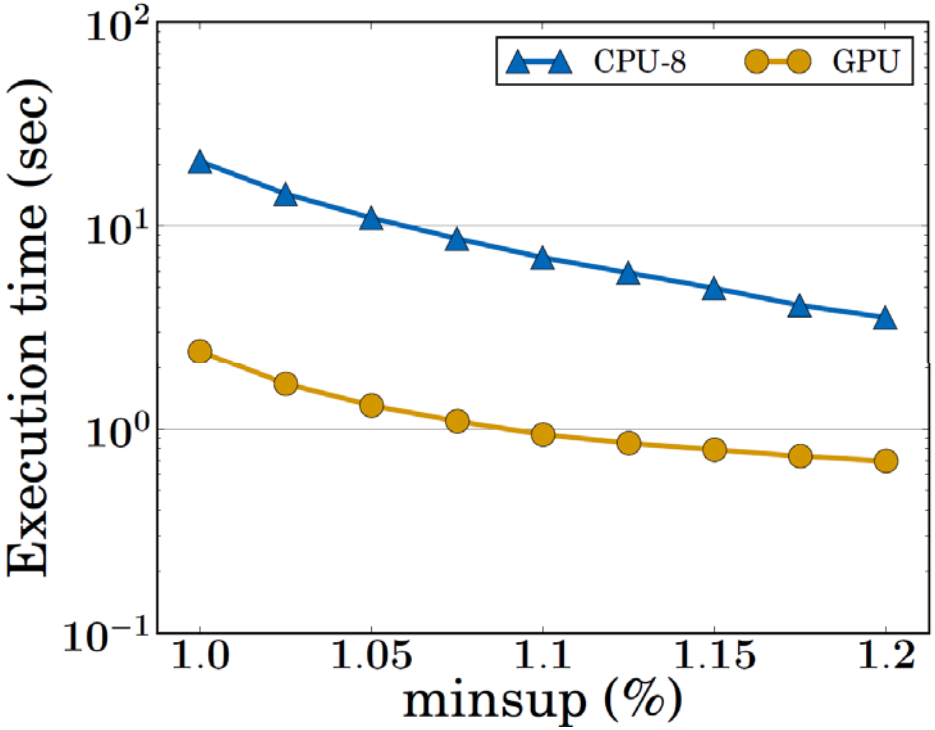
  - Convolution can be efficiently computed with the Fast Fourier Transform (FFT) algorithm
  - Parallelize FFT computation to improve the performance
- Pruning
  - cnt(*X*): the maximum possible value of sup(*X*)
  - esup(*X*): the expected support of *X*

Experimental results

CPU: Intel Xeon (2.40 GHz, 4 cores), 12 GB memory, OpenMP
GPU: Tesla C2050 (1.15 GHz, 448 cores, 3 GB memory), CUDA

- Accidents: 23x–27x
- T25I10D500K: 5.1x–8.6x





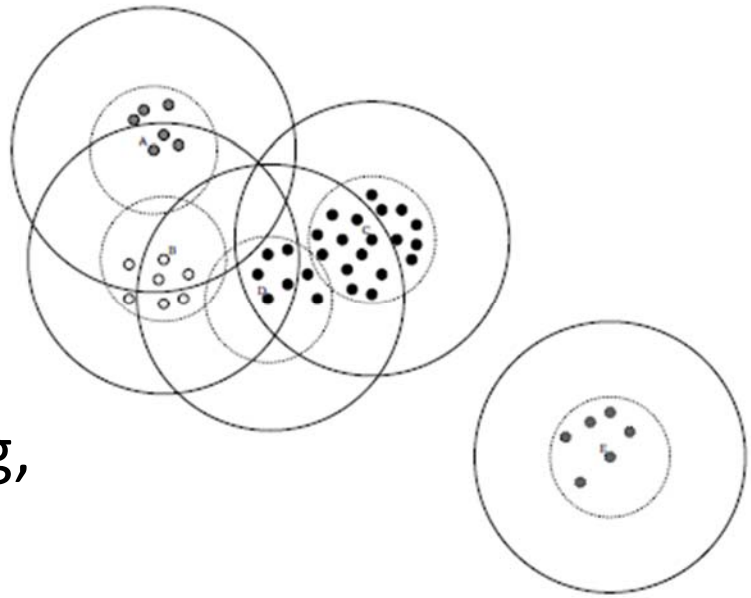
GPU-based Canopy Clustering

Fumitaka Hayashi, Yusuke Kozawa,
Toshiyuki Amagasa, Hiroyuki Kitagawa

University of Tsukuba

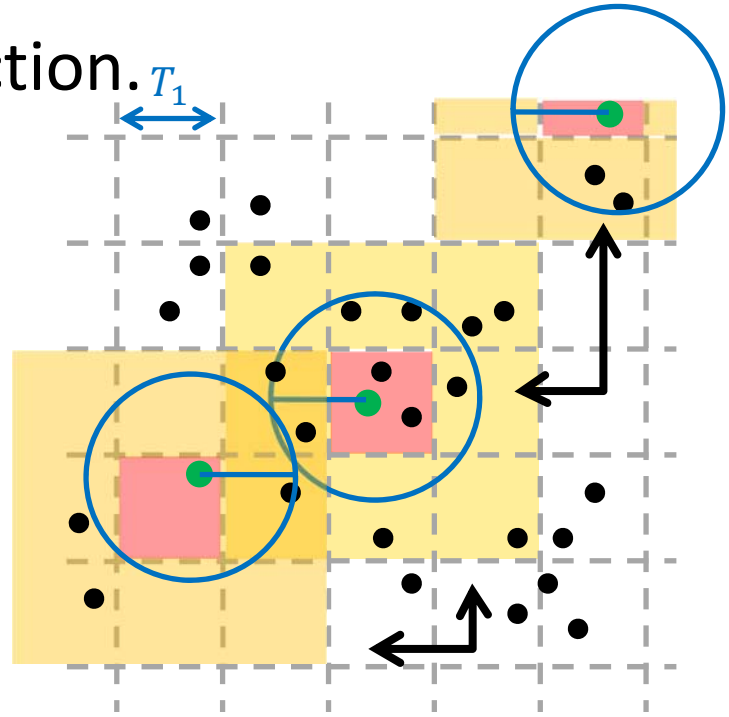
Canopy clustering

- Canopy represents a set of data points.
- Used to speed up clustering algorithms to deal with Big Data.
  - k-means, hierarchical clustering, etc.
  - Tuli De et al. successfully applied clustering to the spectrum of light from extragalactic objects with 700,000 x 1,500 size.

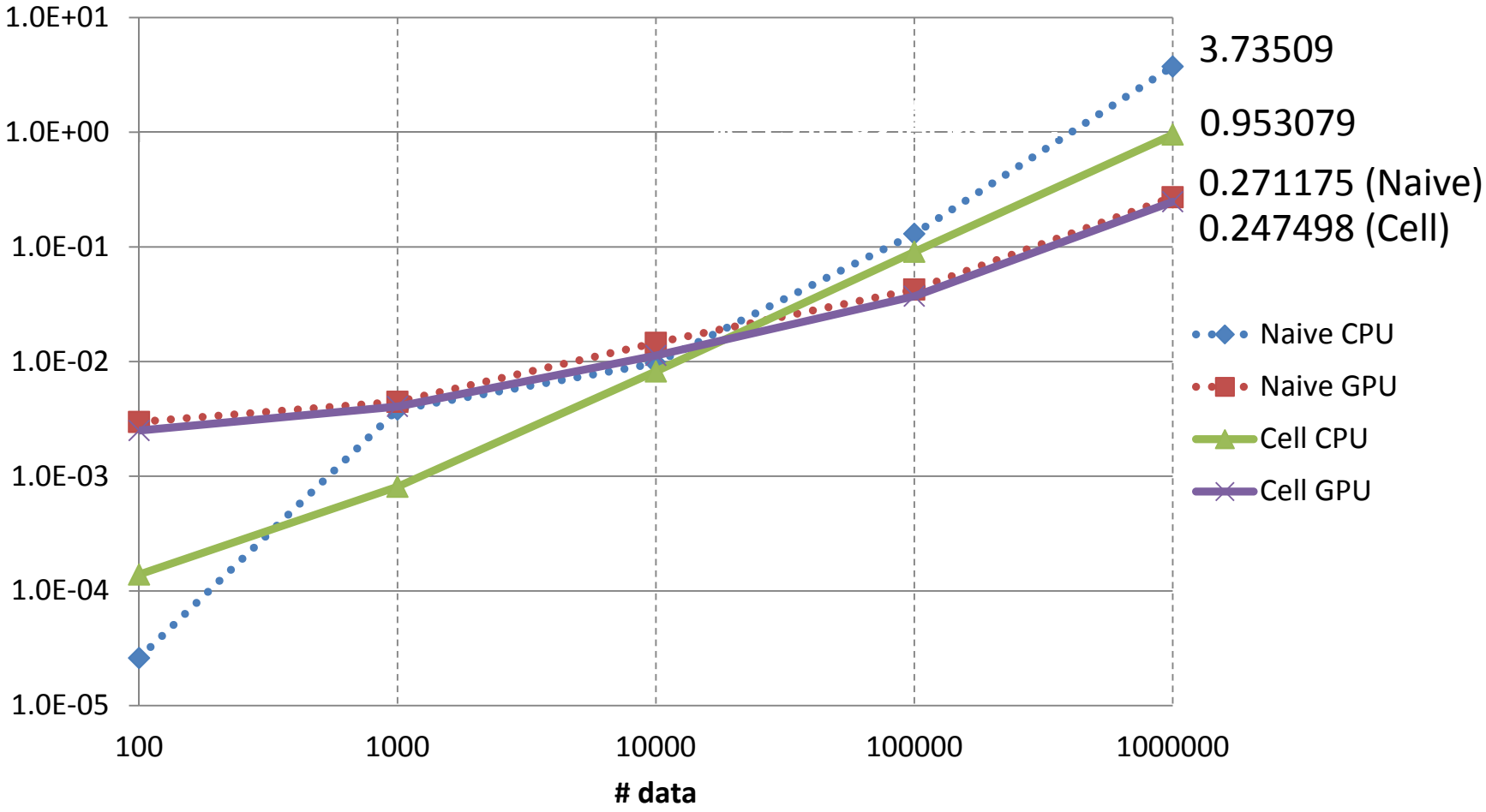


GPU-based canopy clustering Idea

- Parallelize distance computations
  - Each thread computes the distance between a data to the center, and compares with T1 and T2.
  - Intensively use parallel reduction.
- Further optimization
  - Cell-structure
  - Prune unnecessary distance computation



Experimental results

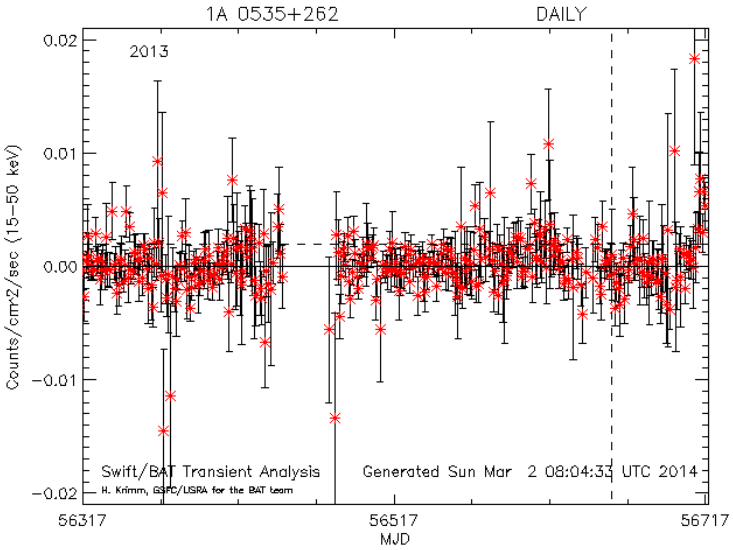


GPU-based Search of Uncertain Time Series

Jun Hwang, Yusuke Kozawa
Toshiyuki Amagasa, Hiroyuki Kitagawa
University of Tsukuba

Uncertain time series search

- Real-world time series often contain uncertainty.
  - e.g., light curve of an X-ray object
- Find similar time series over uncertain time series.
  - MUNICH [Aßfalg et al, 2009]
  - PROUD [Yeh et al, 2009)]
  - **DUST [Sarangi et al, 2010]**



GPU-based acceleration of uncertain time series search

- Idea
  - Parallelize probability computation.
  - DUST
    - Use Monte Carlo integration to compute probability.
    - The performance bottleneck.
- Performance
  - About 230x faster than the naïve CPU-based implementation

Future collaboration

- Improve the performance of data mining over Big Data using GPU/Xeon Phi

- Scientific data management
  - Search over Big Scientific Data
  - Metadata management
    - Linked Open Data
    - XML

Thank you!