

Data Challenges for Data Mining

Philip S. Yu

Outline

- What kind of data should we look for to help move data mining forward?
- Simulated data or real data, what are the pros and cons?
- What can we learn from other empirical sciences in their use of data?

What kind of data is needed

- Real world data is always needed to show the effectiveness of the data mining algorithms
 - Very domain specific
- Even simulated data needs to base on characteristic of real data

Simulated data vs real data

- Simulated data
 - Usually has a data model
 - Provides a better understanding of the causality relation via sensitivity analysis
- Real data
 - Reflect the real world phenomenon
 - More like a black box: Generally it is hard to
 - Model or characterize the real data
 - Explain why the mining techniques work successfully or unsuccessfully

Other Examples

- CPU cache design
 - Cache replacement algorithm
- System performance analysis
 - Prediction of response time and study sensitivity to capacity