

Data mining: lecture 3

Edo liberty

1. Frequent item sets in market basket data
 - (a) problem definition.
 - (b) Sampling works in this case... (just sample a small amount of baskets and compute offline)
 - (c) Trie, prefix tree.
 - (d) Apriory
 - (e) DIC
 - (f) Can we improve DIC by sampling?
 - (g) Can we improve it in any other way?
2. Frequent items in streams
 - (a) Finding an item that is more than half the entries in a stream
 - (b) finding Θ fraction items (A Simple Algorithm for Finding Frequent Elements in Streams and Bags)
 - (c) bloom filters for membership in sets, better than hash functions.
 - (d) Min sketching (Finding Frequent Items in Data Streams)