

Still More Stream-Mining

Frequent Itemsets

Elephants and Troops

Exponentially Decaying Windows

Counting Items

- ◆ **Problem:** given a stream, which items appear more than s times in the window?
- ◆ **Possible solution:** think of the stream of baskets as one binary stream per item.
 - ◆ 1 = item present; 0 = not present.
 - ◆ Use DGIM to estimate counts of 1's for all items.

Extensions

- ◆ In principle, you could count frequent pairs or even larger sets the same way.
 - ◆ One stream per itemset.
- ◆ Drawbacks:
 1. Only approximate.
 2. Number of itemsets is way too big.

Approaches

1. “Elephants and troops”: a heuristic way to converge on unusually strongly connected itemsets.
2. Exponentially decaying windows: a heuristic for selecting likely frequent itemsets.

Elephants and Troops

- ◆ When Sergey Brin wasn't worrying about Google, he tried the following experiment.
- ◆ **Goal:** find unusually correlated sets of words.
 - ◆ "*High Correlation*" = frequency of occurrence of set \gg product of frequency of members.

Experimental Setup

- ◆ The data was an early Google crawl of the Stanford Web.
- ◆ Each night, the data would be streamed to a process that counted a preselected collection of itemsets.
 - ◆ If $\{a, b, c\}$ is selected, count $\{a, b, c\}$, $\{a\}$, $\{b\}$, and $\{c\}$.
 - ◆ "Correlation" = $n^2 * \#abc / (\#a * \#b * \#c)$.
 - n = number of pages.

After Each Night's Processing . . .

1. Find the most correlated sets counted.
2. Construct a new collection of itemsets to count the next night.
 - ◆ All the most correlated sets ("*winners*").
 - ◆ Pairs of a word in some winner and a random word.
 - ◆ Winners combined in various ways.
 - ◆ Some random pairs.

After a Week . . .

- ◆ The pair {"elephants", "troops"} came up as the big winner.
- ◆ **Why?** It turns out that Stanford students were playing a Punic-War simulation game internationally, where moves were sent by Web pages.

Mining Streams Vs. Mining DB's (New Topic)

- ◆ Unlike mining databases, mining streams doesn't have a fixed answer.
- ◆ We're really mining in the "Stat" point of view, e.g., "Which itemsets are frequent in the underlying model that generates the stream?"

Stationarity

- ◆ Two different assumptions make a big difference.
 1. Is the model *stationary*?
 - ◆ I.e., are the same statistics used throughout all time to generate the stream?
 2. Or does the frequency of generating given items or itemsets change over time?

Some Options

- ◆ We could:
 1. Run periodic experiments, like E&T.
 - ◆ Like SON --- itemset is frequent if it is found frequent on any “day.”
 - ◆ Good for stationary statistics.
 2. Frame the problem as finding all frequent itemsets in an “exponentially decaying window.”
 - ◆ Good for nonstationary statistics.

Exponentially Decaying Windows

- ◆ If stream is a_1, a_2, \dots and we are taking the sum of the stream, take the answer at time t to be: $\sum_{i=1,2,\dots,t} a_i e^{-c(t-i)}$.
- ◆ c is a constant, presumably tiny, like 10^{-6} or 10^{-9} .

Example: Counting Items

- ◆ If each a_i is an “item” we can compute the *characteristic function* of each possible item x as an E.D.W.
- ◆ That is: $\sum_{i=1,2,\dots,t} \delta_i e^{-c(t-i)}$, where $\delta_i = 1$ if $a_i = x$, and 0 otherwise.
 - ◆ Call this sum the *weight* of item x .

Counting Items --- (2)

- ◆ Suppose we want to find those items of weight at least $\frac{1}{2}$.
- ◆ **Important property:** sum over all weights is $e^c / (e^c - 1)$ or very close to $1/c$.
- ◆ Thus: at most $2/c$ items have weight at least $\frac{1}{2}$.

Extension to Larger Itemsets*

- ◆ Count (some) itemsets in an E.D.W.
- ◆ When a basket B comes in:
 1. Multiply all counts by $(1-c)$; drop counts $< 1/2$.
 2. If an item in B is uncounted, create new count.
 3. Add 1 to count of any item in B and to any counted itemset contained in B .
 4. Initiate new counts (next slide).

Initiation of New Counts

- ◆ Start a count for an itemset $S \subseteq B$ if every proper subset of S had a count prior to arrival of basket B .
- ◆ **Example:** Start counting $\{i, j\}$ iff both i and j were counted prior to seeing B .
- ◆ **Example:** Start counting $\{i, j, k\}$ iff $\{i, j\}$, $\{i, k\}$, and $\{j, k\}$ were all counted prior to seeing B .

How Many Counts?

- ◆ Counts for single items = $(2/c)$ times the average number of items in a basket.
- ◆ Counts for larger itemsets = ???. But we are conservative about starting counts of large sets.
 - ◆ If we counted every set we saw, one basket of 20 items would initiate 1M counts.