

Presented at the **27th Annual Conference of the Gesellschaft für Klassifikation (GfKI)**, March 12-14, 2003  
Brandenburg University of Technology Cottbus

## **Generating Synthetic Transaction Data for Tuning Usage Mining Algorithms**

Michael Hahsler  
Wirtschaftsuniversität Wien

### **Need for Synthetic Transaction Data**

- Web as a channel for advertising and selling goods
- Automatic services to improve the interface (e.g. recommender systems)
- Complicated algorithms and heuristics

**-> Standardized data sets with known  
characteristics for comparison and tuning**

## The Association Rule Problem

- A database with a set of transactions
- Each transaction contains the items bought by a customer at one visit

$$X \Rightarrow Y$$

- where X and Y are item sets
- and finding X in a transaction means that it is very likely to also find Y in this transaction (controlled by some quality measures)
- $X \cup Y$  is referred to as a "large item set"

## Quest Synthetic Data Generation Code

based on Agrawal and Srikant (1994)

A set of "large itemsets" is generated:

Sizes from Poisson distr. weight from (exp. distr.)

1. item set randomly

the rest using a subset from the previous set (using an exponentially dist. random variable)

1. Size of transactions (Poisson distr.)
2. transaction contains some "large itemsets" using the weight + dropping some items (corruption level)

## Quest Synthetic Data Generation Code

- Generates a structure that contains exactly what association rule algorithms search for!  
(Apriori algorithm)

**Do real data have the same structure?**

**Do Web data have the same structure?**

## Real World Data Sets

- Zheng, Kohavi, Mason (2001)
- Real world data have a different transaction size than the artificial data set
- Performance improvements of new algorithms do not carry over to real world data (Charm, FR-growth, Apriori, Closet)

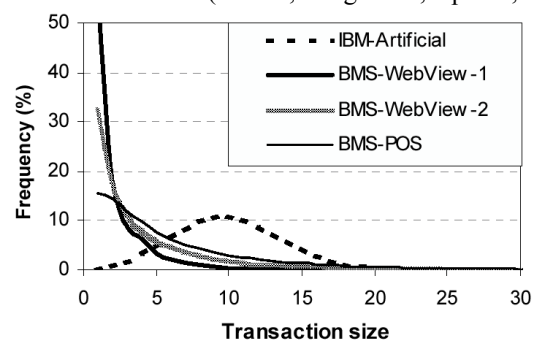
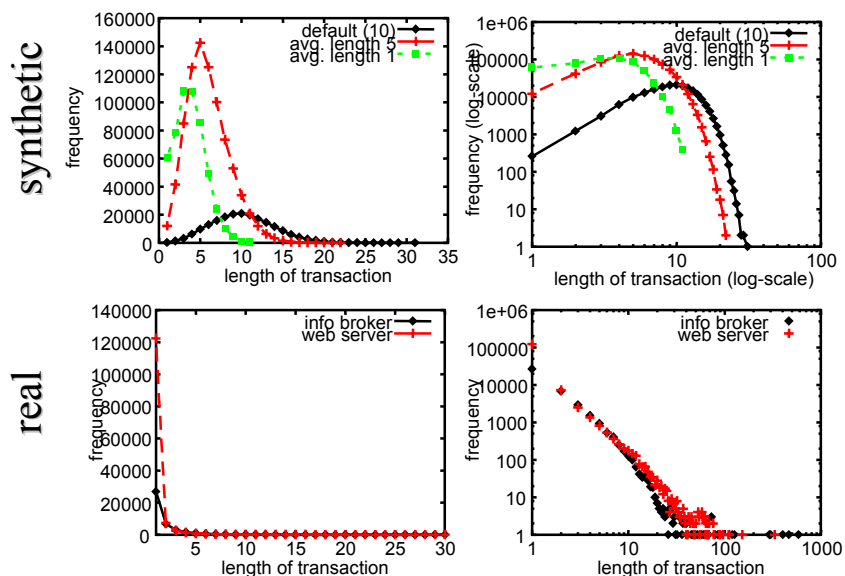


Figure from Zeng et al. (2001)

## Analyze the Characteristics Data Sets

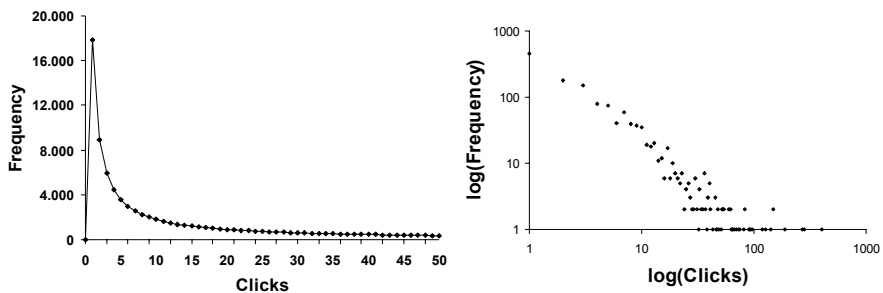
- Several synthetic data sets generated with the Quest generator
- Information broker: A searchable collection of links for students and researchers
- Web server: Preprocessed from the transaction log of the department's Web server

## Transaction Length



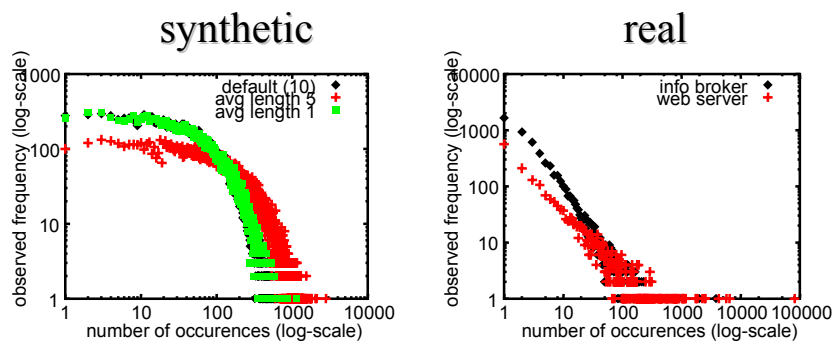
## Pages a user visits within a Web site

- Huberman, Pirolli, Pitkow, Lukose (1998)
- Models page value and stopping.
- The frequency of the number of pages a user visits within a Web site can be modeled as **Inverse Gaussian** distributed
- which is a **Zipf-like** distribution



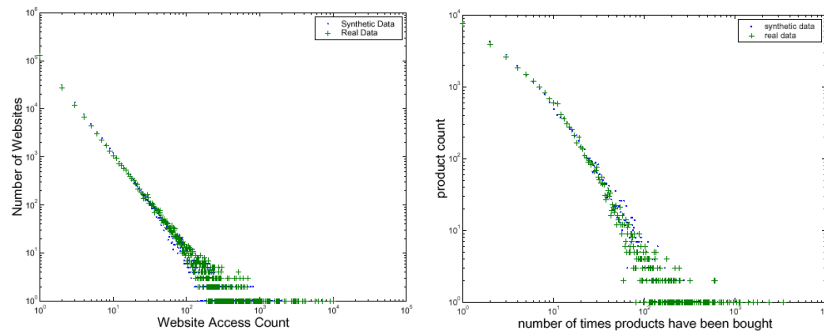
AOL click data from and figures from Huberman et al. (1998)

## Frequency of Different Items



## Frequency of Web Sites by the Number of Visitors

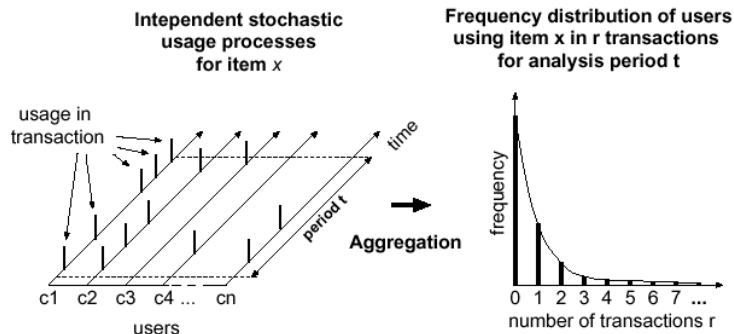
- Bi, Faloutsos, Korn (2001)
- Frequency of Web sites by the number of visitors has a **Zipf-like** distribution
- The count of products by the number of times the product has been bought in a real store can be modeled with a **Discrete Gaussian Exponential** distribution



Figures from Bi et al. (2001)

## The NBD Model

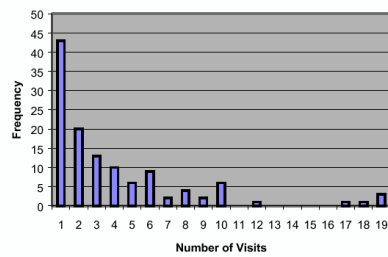
- Model repeat-buying behavior for consumer a good (stationarity)
- Different users ( $c_i$ ) use the item following a Poisson process (the means are drawn from a Gamma dist.)
- The aggregation of all customers leads to a NBD frequency distribution (LSD)



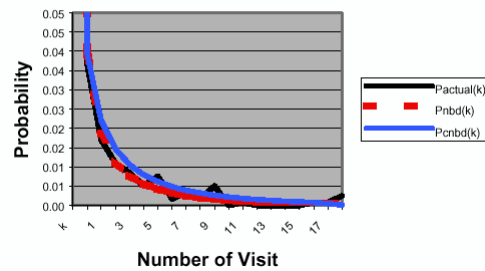
## User Visit Frequency of Web Sites

- Lee, Zufryden, Dreze (2001)
- Model user visit frequency of Web sites using the **Negative Binomial Distribution (NBD)**

Distribution of Visit Frequency  
(Yahoo!)



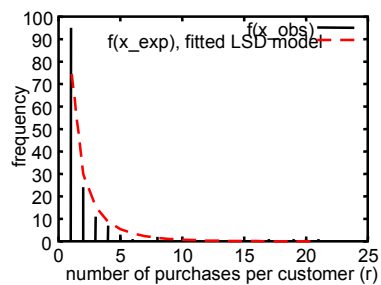
Distribution of Visit Frequency:  
Actual, NBD & CNBD



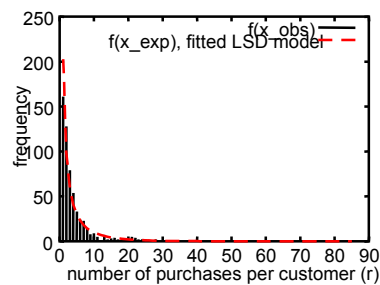
Visits of Yahoo!  
Figures from Lee et al. (2001)

## Usage Frequency by User

Info Broker: External Link  
Alta Vista



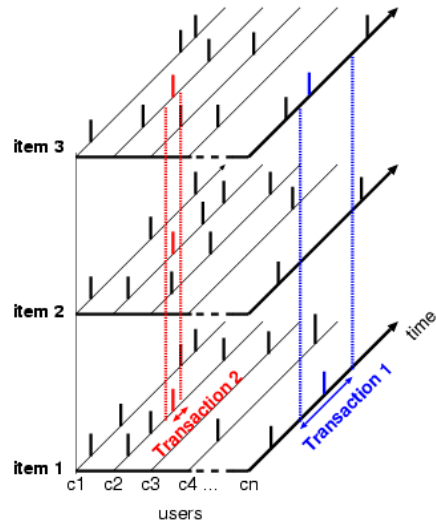
Web Server: Web Pages  
SQL Lecture



The distribution also give a good fit for Web pages  
and information goods

## Generating Synthetic Data using the NBD Model

- We need a generator that creates data sets that have the same characteristics as the real data.
- We simulate the processes described in the NBD model for each item and create transactions



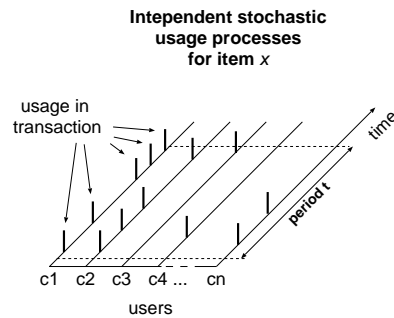
## Generating Synthetic Data using the NBD Model

```
for each item {  
  Initializing the parameters (3) for the Gamma distribution *  
}  
for each user {  
  for each item {  
    draw the parameter for the Poisson process from the item's  
    Gamma distribution  
    produce a list of purchase times (inter-purchase times  
    follow a neg. exponential distribution) **  
  }  
  produce transactions from the purchase times of all items ***  
}
```

\*, \*\* and \*\*\* current research questions

## Current research questions

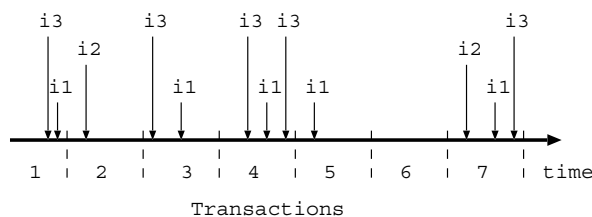
- Distributions for the parameters of the Gamma distributions?
- From real data:



- Estimate the means of the Poisson Processes and then fit a Gamma distribution
- Problem: Not enough observations for most items + stationarity

## Current research questions

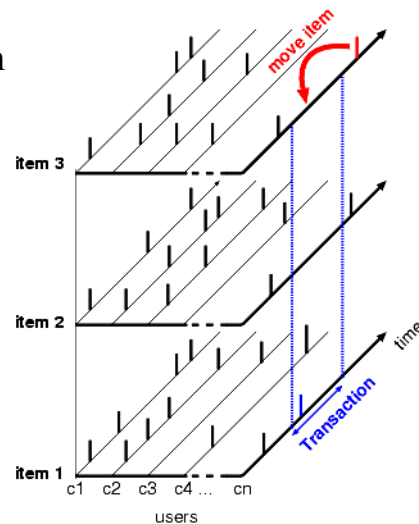
- Transaction length: Distribution?
- Quest uses a Poisson distribution
- Regular intervals



- From the real data sets a Inverse Gaussian distribution of the transaction size seems more appropriate

## Current research questions

- How to incorporate relationships between items and usage patterns?
- Manipulation of the Poisson process by moving a purchase of an item nearer to a related item



## References

- **Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules.** In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487-499, Santiago, Chile, Sept 1994.
- **Andreas Geyer-Schulz, Michael Hahsler, and Maximillian Jahn. A customer purchase incidence model applied to recommender systems.** In R. Kohavi, B.M. Masand, M. Spiliopoulou, and J. Srivastava, editors, *WEBKDD 2001*, LNAI 2356, pages 25-47. Springer-Verlag, July 2002.
- **Bernardo A. Huberman, Peter L. T. Pirolli, James E. Pitkow, and Rajan M. Lukose. Strong regularities in World Wide Web surfing.** *Science*, 280(5360):95-97, 1998.
- **Sukekeyu Lee, Fred Zufryden, and Xavier Dreze. Modeling consumer visit frequency on the internet.** In *34th Annual Hawaii International Conference on System Sciences (HICSS-34)-Volume 7*, 2001.
- **Zijian Zheng, Ron Kohavi, and Llew Mason. Real world performance of association rule algorithms.** In F. Provost and R. Srikant, editors, *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining (ACM-SIGKDD)*, pages 401-406. ACM Press, 2001.
- **Zhiqiang Bi, Christos Faloutsos, and Flip Korn. The "DGX" distribution for mining massive, skewed data.** In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD01)*, pages 17-26, 2001.