

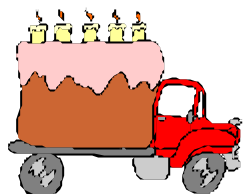
Modeling of Semi-structured Language

Linus Butenas
Vilnius University, 2003

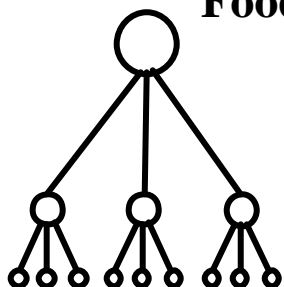
Overview

- Introducing the problem
- The purpose
- Generalized model
- Ideas
- What should be done?

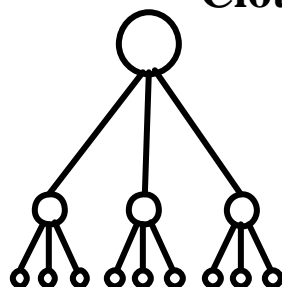
Example



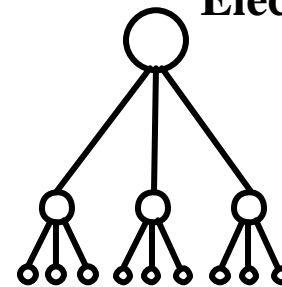
Food



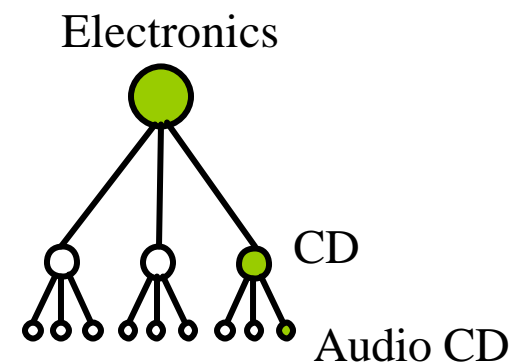
Clothes



Electronics

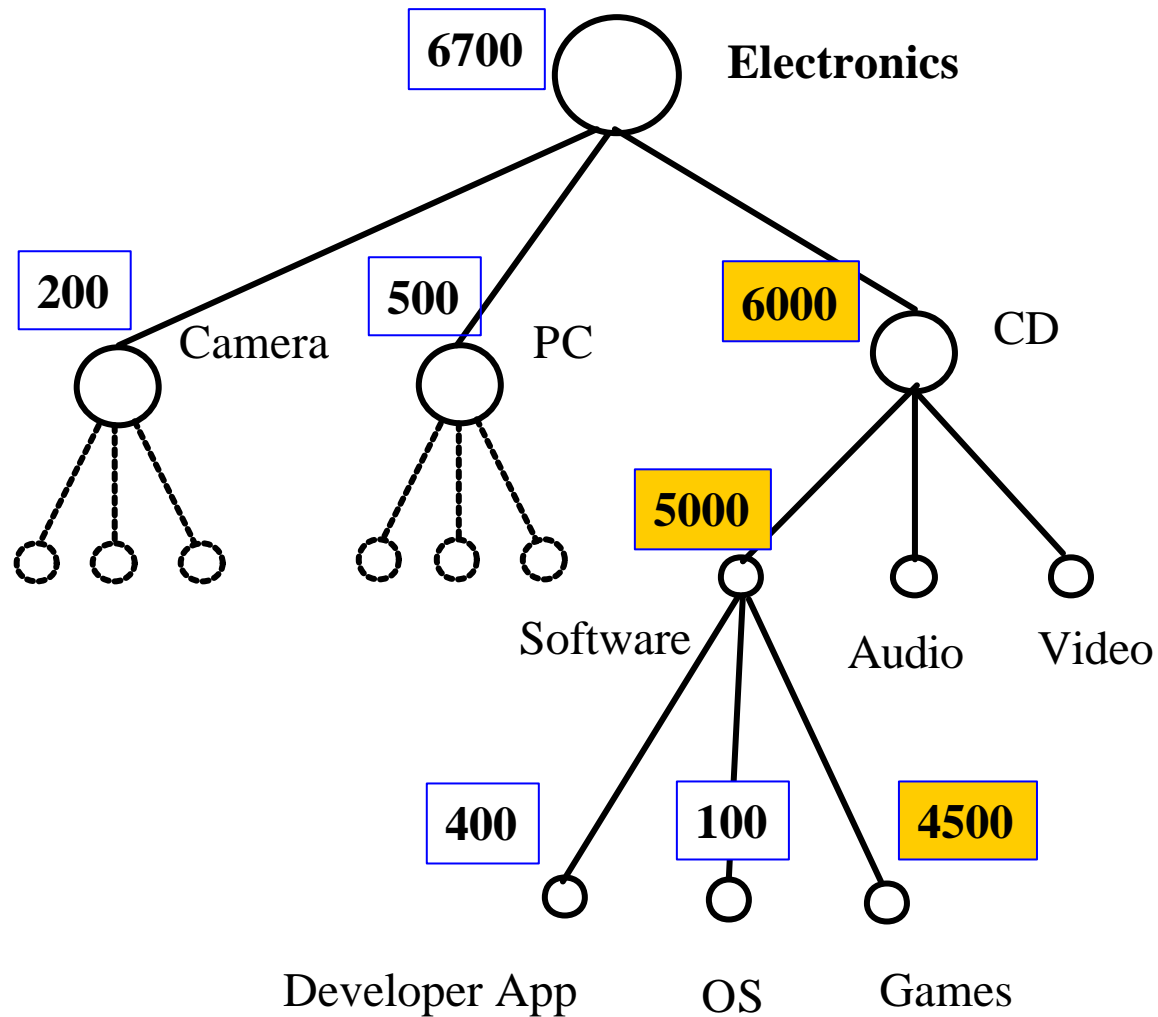


- Each item is described using
 - a *structure* from a category
 - a set of *keywords*



Keywords: music, classic, Bach

Inequality of node use



Keywords indicate new line

- Keywords are a free text

Keywords: Education, Simulation, Action



After some time

Keywords: *PC*, *XBOX*, *PS II*, Education, Action

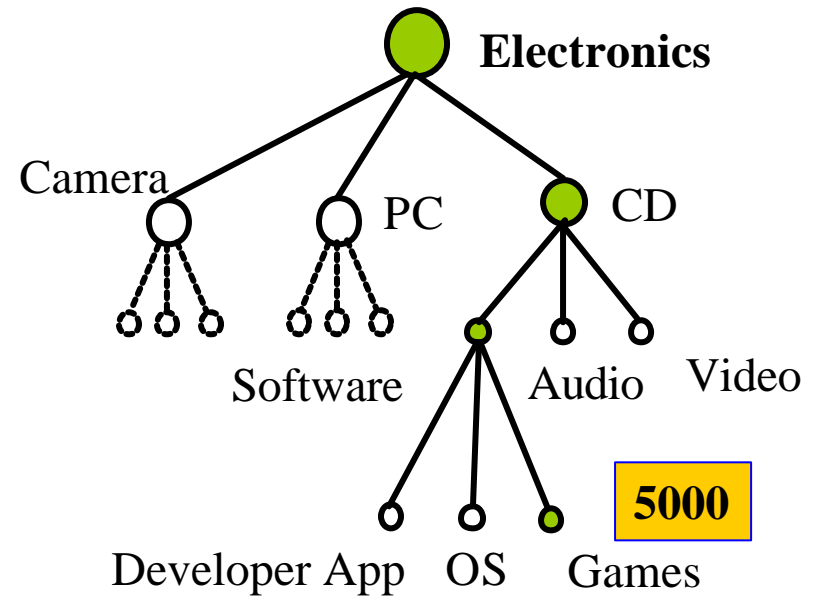
4000

500

500

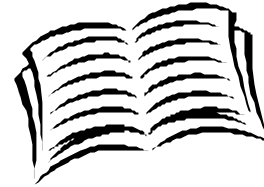
200

3000

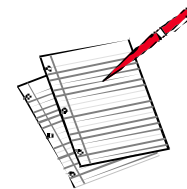


Example with journal

- A mathematical journal

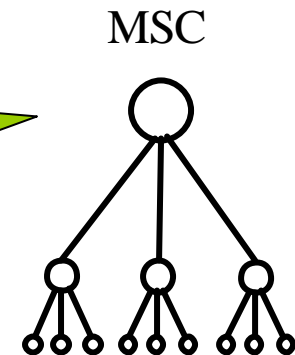
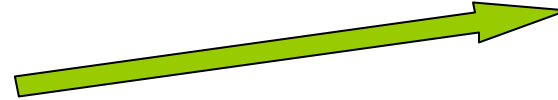


- It consists of papers



- Each paper is described by

- MSC index $xx Y zz$
- Keywords *data mining, ...*



What is the purpose ???

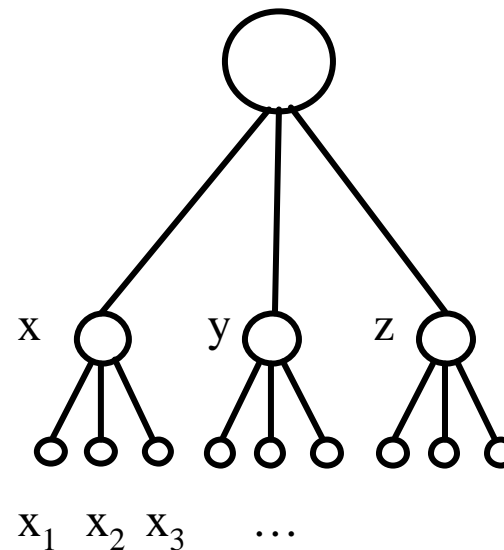
- Journals
 - Splitting the journal to two different journals
- Transport
 - Reorganization of transport infrastructure, reorganization of cargo categorization system
- In general
 - Having a structure, keywords associated with structure and usage of them:
 - estimate and adjust the structure
 - adjust the infrastructure using it

General model

- Usage of nodes and keywords associated to nodes
 - there is no prediction for usage
 - there is a prediction, but after some time the real usage deviates from prediction values

Sets of keywords for nodes:

x_1, x_2, x_3, \dots



Ideas

- Applying statistical methods
 - create a model
 - use statistical methods to determine the changes in a structure
- Viewing a structure as a tree with weights
 - define conditions then a tree structure has to be adjusted
- Using visualization

Data Set and Problems

- A work is based on a small amount of data
 - a number of journals issued during 2-3 years
- Data we have already has no exact cases, showing an exact split of journal

Future work

- Searching for existing models
- Making several new models
- Testing the real life data using models and visualization

Summary

- Maybe we can't make a fully automated model for changing the structure – but we can make a model to notify people

The END

- Questions and ideas are welcome