Powerset Explorer: A Datamining Application

Background

- **PAST**
  - Datamining accomplished with human intuition
- **PRESENT**
  - Computer aided with AI and brute force CPU cycles
- **FUTURE**
  - Enter PowersetViewer...
Dataset
- Alphabet
  - Items that can be found in transactions
  - Eg. Apples, bread, chips
- Transaction
  - Sets of items (unordered)
  - Eg. Tx1 = { Apples, Chips }
  - Eg. Tx2 = { Bread }
- Transaction database
  - Collection of transactions (unordered, possibly repetitive)
  - Eg. Walmart transaction logs

Example Dataset
- Student enrollment database
  - Alphabet = courses
    - { CPSC124, CPSC126, PHIL120, ANTH100, ENGL112 }
**Example Dataset**

- **Student enrollment database**
  - Alphabet = courses
    -  { CPSC124, CPSC126, PHIL120, ANTH100, ENGL112 }
  - Transaction = courses student is enrolled in
    - 239389002 ->  { CPSC 124, PHIL120, ENGL112 }
  - Transaction DB = list of student course schedules

**Example Dataset (cont’d)**

72432398 5 676 1701 3046 3900 1327
38576446 7 175 178 1162 1701 3038 3912
76608235 5 326 676 1701 3038 3908
43539163 3 1177 1699 4327
26495781 6 676 1701 3040 3900
4275

**Why?**

- Why is this interesting?
  - Consumer transaction logs -> trends in consumer buying

**Why? (cont’d)**

- Dataset sizes growing exponentially

---

**Why?**

- Why is this interesting?
  - Consumer transaction logs -> trends in consumer buying
  - Student enrollment database -> trends in enrollment
    - What electives do most undergrad computer science students take?
    - Departments can determine which joint majors would fit the student population.
Why? (cont’d)

- Dataset sizes growing exponentially
  - Human intuition has reached its limits
  - Require computers and AI (expensive)
  - Information visualization can scale the power of human intuition

Powerset Explorer

- Code base from TreeJuxtaposer (Munzner)
  - AccordianDrawer package

Goals
Powerset Explorer

- Code base from TreeJuxtaposer (Munzner)
  - AccordianDrawer package

- Goals
  - Focus + context exploration using grids

Milestones Status Update

- #1 Completion of the basic visualization of a randomized database of small set size (~10)

- #2 Addition of a single level of "marking"
Milestones Status Update

- #1 Completion of the basic visualization of a randomized database of small set size (~10)
- #2 Addition of a single level of “marking”.
- #3 Addition of multiple levels of “marking” (6)
- #4 Addition of background marking to demarcate areas of sets containing different amounts of items.
- #5 Implement multiple constraints
- #6 Increase maximum possible dataset size to at least 100.

Difficulties

- Multiple constraints difficult to implement on current server-side dataminer
Difficulties

- Multiple constraints difficult to implement on current server-side dataminer
- Can not enumerate a powerset of alphabet size greater than 14 elements (integer = 32 bits)
  - Solution: use java class BigInteger

High CPU and memory usage
- Solution: upgrade computer! "hack"

Current Status

- Reduced database
  - 8680433 3 0 7 5
  - 2768129 2 6 4
  - 6385688 5 1 9 10 11
  - 147924 5 5 2 9 5 2
  - 234140 3 1 4 8
  - 4331093 4 4 6 0 0
  - 3158394 5 1 1 1 2 5 4
  - 5797538 6 1 1 3 13 12 4
  - 6243191 1 5
  - 5872060 4 3 8 9 6

- Property file
  - 0 CPSC 325 75.0 3
  - 1 PHIL 327 84.0 1
  - 2 ANTH 329 45.0 2
  - 3 MATH 327 0.0 3
  - 4 PSYC 328 0.0 1
  - 5 ENGL 329 0.0 2
  - 6 APSC 540 0.0 1
  - 7 MENG 541 0.0 1
  - 8 STAT 543 0.0 1
  - 9 SPAN 201 71.0 1
  - 10 FREN 258 76.0 2
  - 11 ECON 260 84.0 1
  - 12 LING 295 42.0 1
  - 13 EECE 302 73.0 1