Powerset Explorer: A Datamining Application

Jordan Lee

1

Background

2

Background

- PAST
  - Datamining accomplished with human intuition

3

Background

- PAST
  - Datamining accomplished with human intuition
- PRESENT
  - Computer aided with AI and brute force CPU cycles

4

Background

- PAST
  - Datamining accomplished with human intuition
- PRESENT
  - Computer aided with AI and brute force CPU cycles
- FUTURE
  - Enter PowersetViewer….

5

Dataset

6

Dataset

- Alphabet
  - Items that can be found in transactions
  - Eg. Apples, bread, chips

7

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 }

8

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

9

Example Dataset

- Student enrollment database

10

Example Dataset

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

11

Example Dataset

- Student enrollment database
  - Alphabet = courses
    - { CPSC124, CPSC126, PHIL120, ANTH100, ENGL112 }
  - Transaction = courses student is enrolled in
    - #29389002 -> { CPSC 124, PHIL120, ENGL112 }

12

Example Dataset

- Student enrollment database
  - Alphabet = courses
    - { CPSC124, CPSC126, PHIL120, ANTH100, ENGL112 }
  - Transaction = courses student is enrolled in
    - #29389002 -> { CPSC 124, PHIL120, ENGL112 }
  - Transaction DB = list of student course schedules

13

Example Dataset (cont'd)

72423298 5 676 1701 3046 3900 1327
38578546 7 175 178 1182 1701 3038 680 3912
7660625 5 326 676 1701 3038 3908
43359163 3 1177 1699 4317
26495781 6 676 1177 1701 3038 3900 4275
48536452 4 1699 2339 1327 2826
64251972 6 676 1177 1701 3038 3900 2549
23212318 5 676 1701 3040 3813 3900
19820119 5 104 676 1699 3038 3900
65954629 4 480 676 3040 3908
54392012 5 676 1701 3038 3813 3899
85833501 5 676 1699 3040 3813 3900
65136197 5 676 1699 3038 3900 2580

14

Why?

- Why is this interesting?

15

Why?

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

16

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.

17

Why? (cont'd)

- Dataset sizes growing exponentially

18

Why? (cont'd)

- Dataset sizes growing exponentially
  - Human intuition has reached its limits

19

Why? (cont'd)

- Dataset sizes growing exponentially
  - Human intuition has reached its limits
  - Require computers and AI (expensive)

20

Why? (cont'd)

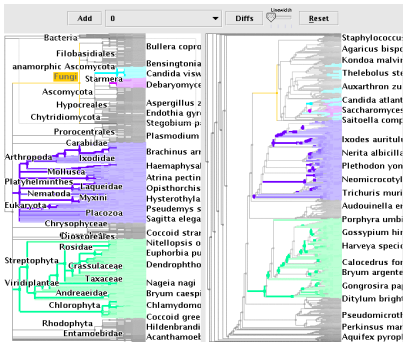
- Dataset sizes growing exponentially
  - Human intuition has reached its limits
  - Require computers and AI (expensive)
  - Information visualization can scale the power of human intuition

21

Powerset Explorer

- Code base from TreeJuxtaposer (Munzner)
  - AccordianDrawer package

22



TreeJuxtaposer

Powerset Explorer

- Code base from TreeJuxtaposer (Munzner)
  - AccordianDrawer package
- Goals

24

Powerset Explorer

- Code base from TreeJuxtaposer (Munzner)
  - AccordianDrawer package
- Goals
  - Focus + context exploration using grids

25

Powerset Explorer

- Code base from TreeJuxtaposer (Munzner)
  - AccordianDrawer package
- Goals
  - Focus + context exploration using grids
  - Guaranteed visibility

26

Powerset Explorer

- Code base from TreeJuxtaposer (Munzner)
  - AccordianDrawer package
- Goals
  - Focus + context exploration using grids
  - Guaranteed visibility
  - Marking of groups

27

Milestones Status Update

28

Milestones Status Update

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

29

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”.

30

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)

31

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.

32

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

33

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.

34

Difficulties

35

Difficulties

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

36

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

37

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
  - Solultion: upgrade computer! ←hack

38

Current Status

- Reduced database

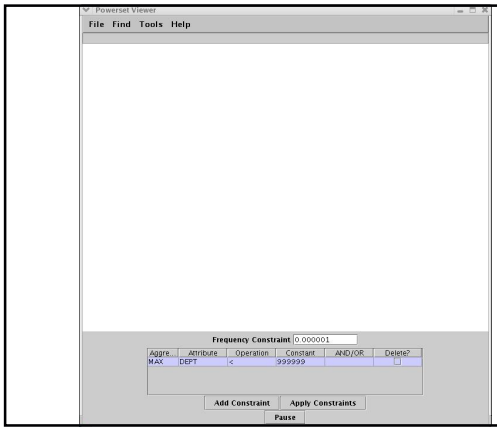
8680433 3 0 7 5
2768129 2 6 4
6385608 5 1 9 10 9 11
147924 5 5 2 9 5 2
234140 3 11 4 8
4331093 4 4 6 0 0
3158394 5 12 1 12 5 4
5797538 6 11 4 3 13 12 4
6243191 1 5
5872060 4 3 8 9 6

39

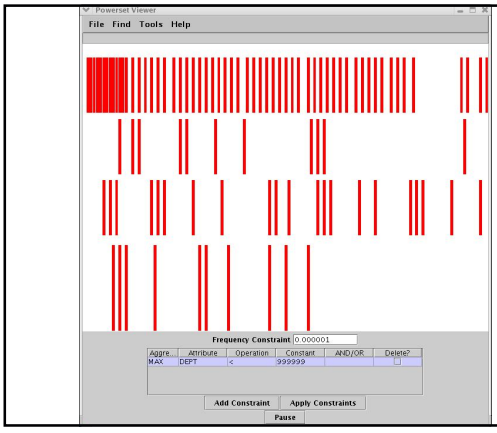
Current Status

- 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 MECH 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

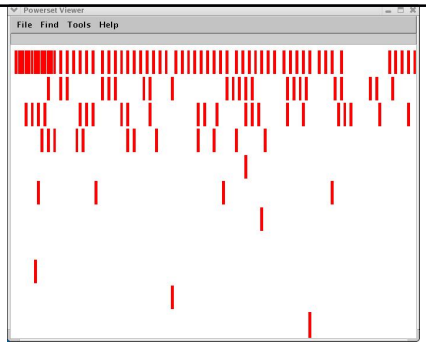
40



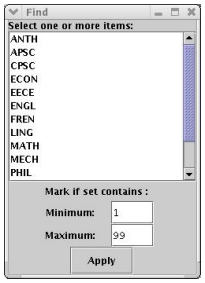
41



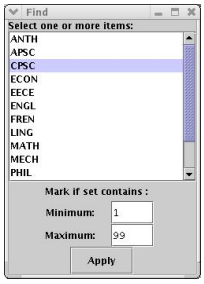
42



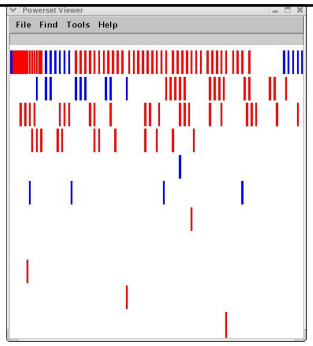
43



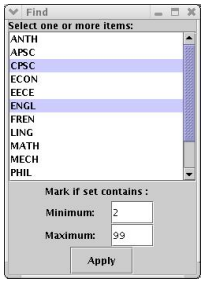
44



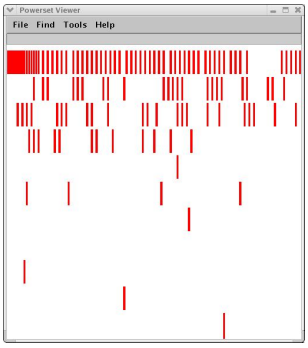
45



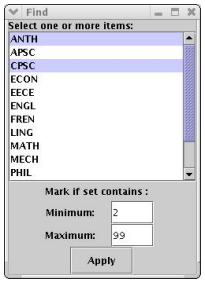
46



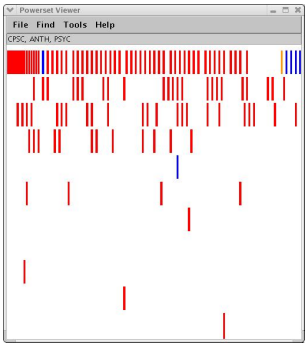
47



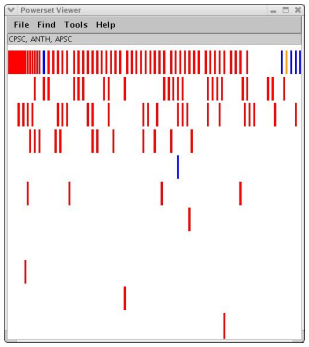
48



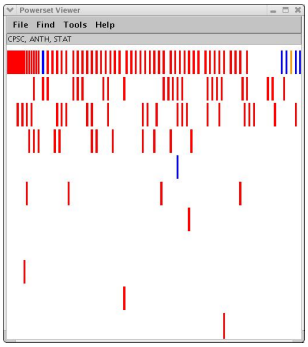
49



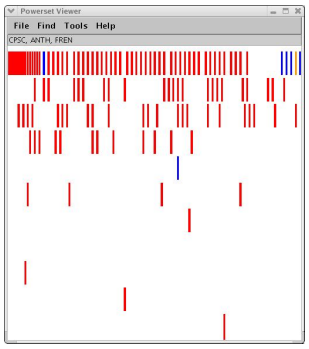
50



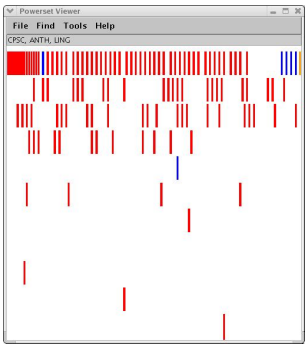
51



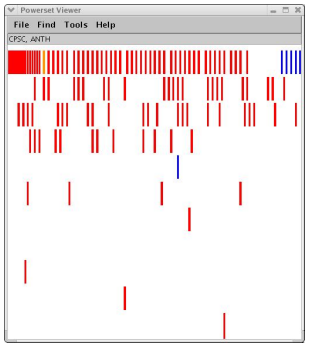
52



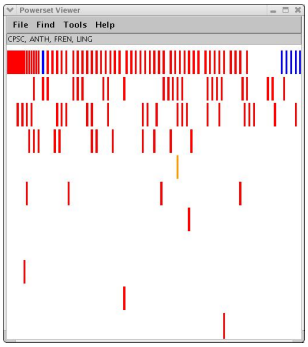
53



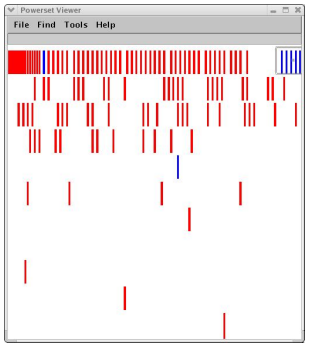
54



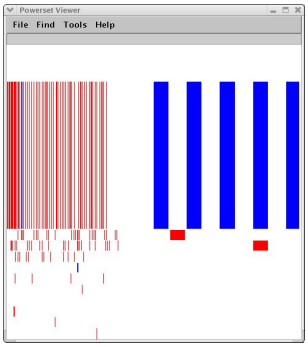
55



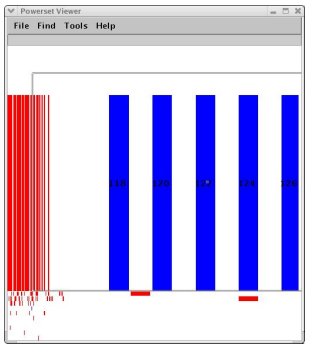
56



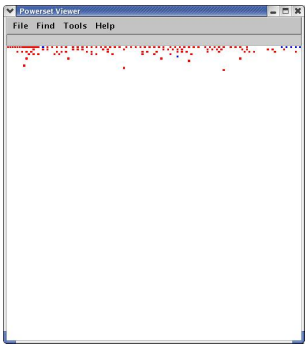
57



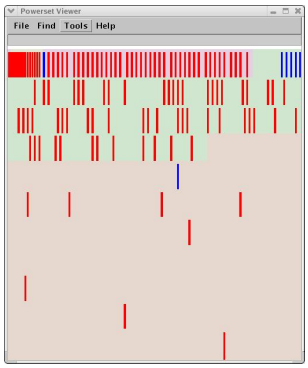
58



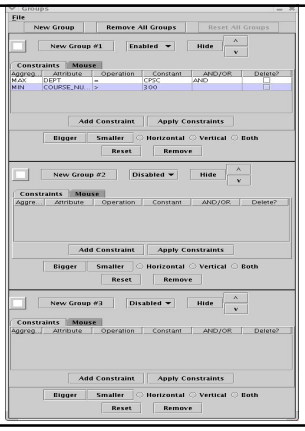
59



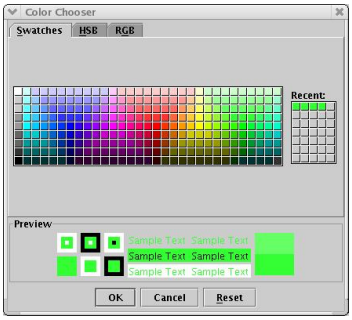
60



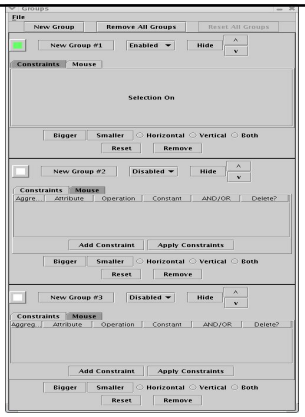
61



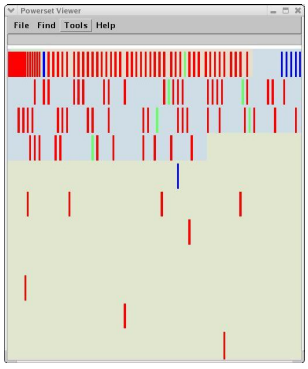
62



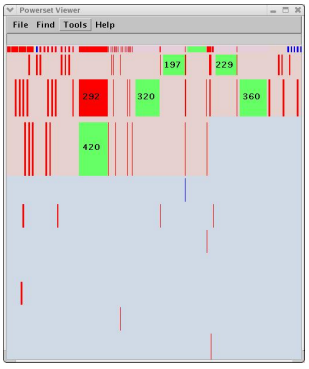
63



64



65



66

Questions?

67