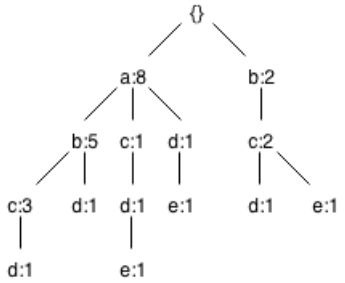
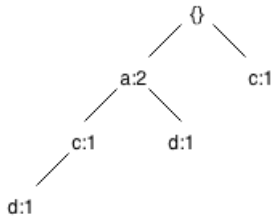
**CS6220 Data Mining Fall 2014 Homework 3, Wei Luo**

**1. Frequent Pattern Mining for Set Data**

(a) Scan the Database once, we get:
a:8 b:7 c:6 d:5 e:3
Sort them and build the FP-tree:




(b) e's conditional pattern base is:
acd:1, ad:1, bc:1
So for e's conditional FP-tree, we have:
a:2 b:1 c:2 d:2 , remove b:1 since it doesn't reach the min_support=2, e's conditional FP-tree is:




Frequent patterns based on e's conditional FP-tree: e, ae, ce, de, ade, cde, ace, acde.

**2. Correction Analysis**

(a) Based on the observed values given, we can calculate the probabilities and get the observed values table:

| | Beer | No Beer | Total |
|---|---|---|---|
| Nuts | 17 | 833 | 850 |
| No Nuts | 183 | 8967 | 9150 |
| Total | 200 | 9800 | 10000 |

$confidence(Beer \Rightarrow Nuts) = 50/200 = 0.25$
$lift(Beer, Nuts) = \frac{P(Beer \cup Nuts)}{P(Beer)P(Nuts)} = \frac{50/10000}{200/10000*850/10000} = 2.9412$
$\chi^2 = \frac{(50-17)^2}{17} + \frac{(800-833)^2}{833} + \frac{(150-183)^2}{183} + \frac{(9000-8967)^2}{8967} = 71.4384$
$all\_confidence = min(P(Beer|Nuts), P(Nuts|Beer)) = min(50/850, 50/200) = 0.0588$
(b) Since $lift(Beer, Nuts) = 2.9412 > 1$. Buying beer and buying nuts are positively correlated.

**3. Sequential Pattern Mining (GSP Algorithm)**

(a) For a sequence $s = \langle (ab)(cd)ef \rangle$. $s$ contains 4 elements. The length of $s$ is 8. It contains 63 non-empty subsequences.

(b) From $L_3$ drop the first and last element of each sequence, we can get:

| ID | s | drop first | drop last |
|---|---|---|---|
| 1. | $\langle (ab)c \rangle$ | $\langle (b)c \rangle$ | $\langle (ab) \rangle$ |
| 2. | $\langle (ab)d \rangle$ | $\langle (b)d \rangle$ | $\langle (ab) \rangle$ |
| 3. | $\langle a(cd) \rangle$ | $\langle (cd) \rangle$ | $\langle a(c) \rangle$ |
| 4. | $\langle (ac)e \rangle$ | $\langle (c)e \rangle$ | $\langle (ac) \rangle$ |
| 5. | $\langle b(cd) \rangle$ | $\langle (cd) \rangle$ | $\langle b(c) \rangle$ |
| 6. | $\langle bce \rangle$ | $\langle ce \rangle$ | $\langle bc \rangle$ |

We can see that 1 and 5 , 1 and 6 can be joined. Join them we get:
$\langle (ab)(cd) \rangle$, $\langle (ab)ce \rangle$
For $\langle (ab)(cd) \rangle$, all its length 3 subsequences are in $L_3$, keep it.
For $\langle (ab)ce \rangle$, its subsequence $\langle (ab)e \rangle$ is not in $L_3$, so we prune it.
So $C_4$ is $\langle (ab)(cd) \rangle$.

**4. Application**

First, get the titles of papers in the 20 conferences in time periods 2001-2005, 2008-2012. Then, use tools from nltk to get tokens from each title. Ignore stop words, punctuations and numbers. After that, use PyFIM with APRIORI algorithm to find the frequent sequence patterns.
The results are:

For year 2001-2005:
Top 20 most frequent patterns with length 1

| | | | |
|---|---|---|---|
| ⟨data⟩ :1231 | ⟨information⟩ :418 | ⟨efficient⟩ :289 | ⟨queries⟩ :226 |
| ⟨based⟩ :764 | ⟨xml⟩ :383 | ⟨approach⟩ :276 | ⟨systems⟩ :225 |
| ⟨mining⟩ :665 | ⟨retrieval⟩ :316 | ⟨classification⟩ :264 | |
| ⟨using⟩ :607 | ⟨search⟩ :314 | ⟨system⟩ :243 | |
| ⟨web⟩ :591 | ⟨query⟩ :313 | ⟨database⟩ :238 | |
| ⟨learning⟩ :521 | ⟨clustering⟩ :307 | ⟨model⟩ :231 | |

Top 20 most frequent patterns with length 2

| | | |
|---|---|---|
| ⟨mining data⟩ :297 | ⟨web data⟩ :74 | ⟨association mining⟩ :55 |
| ⟨based data⟩ :105 | ⟨web mining⟩ :74 | ⟨learning data⟩ :54 |
| ⟨streams data⟩ :95 | ⟨management data⟩ :72 | ⟨efficient data⟩ :53 |
| ⟨using data⟩ :88 | ⟨web based⟩ :70 | ⟨model based⟩ :53 |
| ⟨clustering data⟩ :82 | ⟨frequent mining⟩ :67 | ⟨dimensional data⟩ :51 |
| ⟨xml data⟩ :82 | ⟨patterns mining⟩ :62 | ⟨learning based⟩ :50 |
| ⟨clustering based⟩ :75 | ⟨approach based⟩ :58 | |

Top 20 most frequent patterns with length 3

| | | |
|---|---|---|
| ⟨asia mining data⟩ :6 | ⟨cliques data mining⟩ :2 | ⟨ssp mining data⟩ :2 |
| ⟨mdm mining data⟩ :6 | ⟨drifting mining data⟩ :2 | ⟨ugly mining data⟩ :2 |
| ⟨pacific mining data⟩ :6 | ⟨fractals mining data⟩ :2 | ⟨warehouse based data⟩ :2 |
| ⟨dm mining data⟩ :3 | ⟨grids mining data⟩ :2 | ⟨7th data mining⟩ :1 |
| ⟨pakdd mining data⟩ :3 | ⟨medicine mining data⟩ :2 | ⟨aaand mining using⟩ :1 |
| ⟨academy mining data⟩ :2 | ⟨ole mining data⟩ :2 | ⟨aboutness using based⟩ :1 |
| ⟨bad mining data⟩ :2 | ⟨peculiarity data mining⟩ :2 | |

Top 20 most frequent patterns with length 4

⟨adherence using based data⟩ :1
⟨admit mining based data⟩ :1
⟨bibfinder mining using data⟩ :1
⟨columbia mining based data⟩ :1
⟨cooperatively mining using data⟩ :1
⟨deployment mining based data⟩ :1
⟨divisive mining using data⟩ :1
⟨effectively mining using data⟩ :1
⟨gdt mining using data⟩ :1
⟨generalised mining using data⟩ :1
⟨ibl mining using data⟩ :1
⟨ids mining based data⟩ :1
⟨infer mining using data⟩ :1
⟨lead mining using data⟩ :1
⟨meningitis mining using data⟩ :1
⟨rs mining using data⟩ :1
⟨rsbr mining using data⟩ :1
⟨simplicial mining using data⟩ :1
⟨statminer mining using data⟩ :1
⟨ubdm mining based data⟩ :1

For year 2008-2012:
Top 20 most frequent patterns with length 1

⟨data⟩ :1856
⟨based⟩ :1783
⟨using⟩ :1129
⟨learning⟩ :1099
⟨mining⟩ :1004
⟨search⟩ :738
⟨information⟩ :637
⟨analysis⟩ :617
⟨web⟩ :596
⟨system⟩ :569
⟨classification⟩ :546
⟨query⟩ :542
⟨efficient⟩ :511
⟨clustering⟩ :505
⟨retrieval⟩ :493
⟨model⟩ :484
⟨networks⟩ :480
⟨approach⟩ :474
⟨multi⟩ :474
⟨time⟩ :386

Top 20 most frequent patterns with length 2

⟨mining data⟩ :421
⟨data based⟩ :179
⟨system based⟩ :152
⟨model based⟩ :136
⟨streams data⟩ :127
⟨using data⟩ :127
⟨approach based⟩ :123
⟨machine learning⟩ :123
⟨clustering based⟩ :119
⟨using based⟩ :117
⟨learning using⟩ :115
⟨algorithm based⟩ :113
⟨mining based⟩ :111
⟨analysis data⟩ :110
⟨learning based⟩ :108
⟨research based⟩ :100
⟨analysis based⟩ :95
⟨learning data⟩ :94
⟨multi learning⟩ :94
⟨management data⟩ :93

Top 20 most frequent patterns with length 3

⟨warehouses based data⟩ :3
⟨modis using data⟩ :2
⟨pathway data based⟩ :2
⟨abnormalities learning data⟩ :1
⟨abnormalities learning using⟩ :1
⟨abnormalities using data⟩ :1
⟨abundant learning based⟩ :1
⟨accents learning using⟩ :1
⟨aco using based⟩ :1
⟨adverse using data⟩ :1
⟨affected using data⟩ :1
⟨albatross using data⟩ :1
⟨ale using based⟩ :1
⟨alias learning based⟩ :1
⟨allow based data⟩ :1
⟨alphabets based using⟩ :1
⟨alpos learning data⟩ :1
⟨american learning using⟩ :1
⟨analyst data based⟩ :1
⟨ancheng data based⟩ :1

Top 20 most frequent patterns with length 4

⟨abnormalities learning using data⟩ :1
⟨bee using data based⟩ :1
⟨ciphertext using data based⟩ :1
⟨comet learning using data⟩ :1
⟨froc learning using data⟩ :1
⟨hazards using data based⟩ :1
⟨homogenous learning using data⟩ :1
⟨pathway using data based⟩ :1
⟨periodical learning data based⟩ :1
⟨recipe learning using data⟩ :1
⟨remaining using data based⟩ :1
⟨reverible learning using data⟩ :1
⟨rotation learning data based⟩ :1
⟨sigma using data based⟩ :1
⟨subdivision learning data based⟩ :1
⟨tailed using data based⟩ :1
⟨topographic learning using data⟩ :1
⟨vibratory using data based⟩ :1