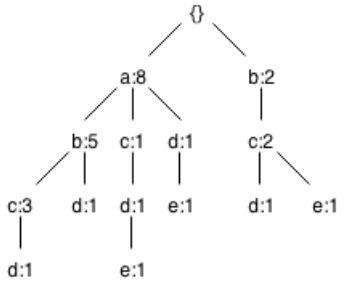
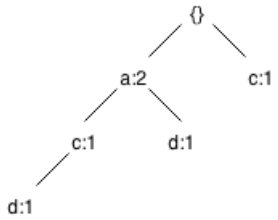
CS6220 Data Mining Fall 2014 Homework 3, Wei Luo

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.

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 6. 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 PrefixSpan 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

⟨data mining⟩ :297	⟨xml data⟩ :82	⟨search web⟩ :72
⟨information retrieval⟩ :151	⟨based clustering⟩ :75	⟨web based⟩ :70
⟨based data⟩ :105	⟨proceedings conference⟩ :74	⟨dimensional high⟩ :69
⟨data streams⟩ :95	⟨web data⟩ :74	⟨frequent mining⟩ :67
⟨data using⟩ :88	⟨web mining⟩ :74	⟨report workshop⟩ :66
⟨time series⟩ :83	⟨association rules⟩ :72	⟨processing query⟩ :64
⟨data clustering⟩ :82	⟨management data⟩ :72	

Top 20 most frequent patterns with length 3

⟨proceedings international conference⟩ :47	⟨association rules mining⟩ :25
⟨high dimensional data⟩ :39	⟨data mining workshop⟩ :25
⟨support vector machines⟩ :39	⟨discovery knowledge data⟩ :24
⟨proceedings conference data⟩ :32	⟨proceedings international data⟩ :24
⟨language information retrieval⟩ :28	⟨proceedings data mining⟩ :24
⟨data mining knowledge⟩ :26	⟨international conference data⟩ :23

⟨proceedings conference mining⟩ :23
⟨report workshop data⟩ :23
⟨rule association mining⟩ :23
⟨conference data mining⟩ :22
⟨discovery knowledge mining⟩ :22
⟨preserving privacy data⟩ :22
⟨workshop retrieval information⟩ :22
⟨series time data⟩ :21

Top 20 most frequent patterns with length 4

⟨proceedings international conference data⟩ :22
⟨proceedings conference data mining⟩ :21
⟨discovery knowledge data mining⟩ :19
⟨proceedings conference discovery knowledge⟩ :18
⟨proceedings discovery knowledge mining⟩ :18
⟨conference discovery mining knowledge⟩ :17
⟨proceedings conference knowledge mining⟩ :17
⟨proceedings knowledge data mining⟩ :17
⟨conference knowledge data mining⟩ :16
⟨proceedings conference discovery mining⟩ :16
⟨proceedings discovery knowledge data⟩ :16
⟨proceedings discovery data mining⟩ :16
⟨conference discovery data knowledge⟩ :15
⟨conference discovery data mining⟩ :15
⟨proceedings conference knowledge data⟩ :15
⟨acm proceedings international conference⟩ :14
⟨cross language retrieval information⟩ :14
⟨dimensional high clustering data⟩ :14
⟨proceedings international conference mining⟩ :14
⟨proceedings conference discovery data⟩ :14

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

⟨data mining⟩ :421
⟨information retrieval⟩ :233
⟨based data⟩ :179
⟨social networks⟩ :155
⟨system based⟩ :152
⟨scale large⟩ :139
⟨model based⟩ :136
⟨feature selection⟩ :131
⟨web search⟩ :128
⟨data streams⟩ :127
⟨using data⟩ :127
⟨approach based⟩ :123
⟨machine learning⟩ :123
⟨time series⟩ :123
⟨semi supervised⟩ :122
⟨clustering based⟩ :119
⟨using based⟩ :117
⟨learning using⟩ :115
⟨algorithm based⟩ :113
⟨mining based⟩ :111

Top 20 most frequent patterns with length 3

⟨proceedings international conference⟩ :63
⟨proceedings data mining⟩ :61
⟨international data mining⟩ :53
⟨proceedings international data⟩ :53
⟨proceedings conference data⟩ :52
⟨semi supervised learning⟩ :51
⟨data mining workshop⟩ :51
⟨conference data mining⟩ :45
⟨high dimensional data⟩ :43
⟨discovery knowledge data⟩ :43
⟨proceedings conference mining⟩ :42
⟨proceedings international mining⟩ :42
⟨international workshop data⟩ :42
⟨data mining based⟩ :42
⟨proceedings discovery knowledge⟩ :41
⟨machines vector support⟩ :38

⟨conference discovery knowledge⟩ :37
⟨international mining workshop⟩ :35
⟨conference international data⟩ :34
⟨series time data⟩ :34

Top 20 most frequent patterns with length 4

⟨proceedings conference data mining⟩ :41
⟨proceedings international data mining⟩ :38
⟨proceedings conference international data⟩ :32
⟨proceedings conference discovery knowledge⟩ :32
⟨discovery knowledge data mining⟩ :27
⟨international workshop data mining⟩ :27
⟨proceedings discovery knowledge data⟩ :25
⟨conference international data mining⟩ :24
⟨proceedings conference international mining⟩ :23
⟨proceedings conference machine learning⟩ :22
⟨part conference discovery knowledge⟩ :22
⟨preface workshop data mining⟩ :22
⟨conference discovery data knowledge⟩ :21
⟨proceedings international workshop data⟩ :21
⟨proceedings discovery knowledge mining⟩ :21
⟨proceedings discovery data mining⟩ :21
⟨conference discovery data mining⟩ :20
⟨conference discovery mining knowledge⟩ :20
⟨conference knowledge data mining⟩ :20
⟨part proceedings conference discovery⟩ :20