CS524

Data Mining

Lab Assignment 2

PART 1

Frequent pattern mining using:
Apriori Algorithm
And
ECLAT Algorithm

Submitted By:
Paras Kumar
2016CSB1047

TASK 1

Apriori Algorithm

I used a hash based technique as described in the book to speed normal Apriori algorithm.

Earlier I was not able to find frequent itemsets of length larger than 2, but, after the optimization it could find itemsets of length upto 4. (Now it supports minimum support >= 0.015%)

Running Time

5.418997049331665 seconds

Runtime configurations

Min support = 0.015%
Intel i5 7th gen CPU with 8GB of RAM

TASK 1

ECLAT Algorithm

I have implemented ECLAT algorithm as described in the book. I first transformed the dataset to take advantage of the vertical data format and then used basic set operations to find the frequent itemsets. This algorithm performs much better than the optimized Apriori algorithm and it can support even lower min support values, about 0.01 % which corresponds to only 2 occurrences of itemsets and itemsets of length 7 can be found in about 20 sec which was not even possible with the previous algorithm.

Running Time

3.360443353652954 seconds

Runtime configurations

Min support = 0.015%
Intel i5 7th gen CPU with 8GB of RAM

TASK 2

For the support value of 0.015% many itemsets of length 2 were generated, so I am not displaying them here, you can find them in my output.txt file.

Sample Output for Apriori algorithm

```
Frequent itemsets of length  3 are:
Mahmut T. Kandemir, J. Ramanujam, Prithviraj Banerjee,
Diego Calvanese, Maurizio Lenzerini, Giuseppe De Giacomo,
Martin Ester, Hans-Peter Kriegel, Jörg Sander,
Germán Vidal, María Alpuente, Moreno Falaschi,
Wolfram Burgard, Sebastian Thrun, Dieter Fox,
Martin Ester, Hans-Peter Kriegel, Xiaowei Xu,
Massimo Violante, Matteo Sonza Reorda, Maurizio Rebaudengo,
Yervant Zorian, Michel Renovell, Joan Figueras,
Yervant Zorian, J. M. Portal, Michel Renovell,
Yervant Zorian, J. M. Portal, Joan Figueras,
J. M. Portal, Michel Renovell, Joan Figueras,
Mahmut T. Kandemir, J. Ramanujam, Alok N. Choudhary,
Mahmut T. Kandemir, Prithviraj Banerjee, Alok N. Choudhary,
Prithviraj Banerjee, J. Ramanujam, Alok N. Choudhary,
Keshav Pingali, Vladimir Kotlyar, Paul Stodghill,
Roland Wismüller, Wlodzimierz Funika, Marian Bubak,
Lipyeow Lim, Jeffrey Scott Vitter, Min Wang,
Henri Prade, Didier Dubois, Salem Benferhat,
Joan Martí, Jordi Freixenet, Xavier Cufí,
Hakan Ferhatosmanoglu, Amr El Abbadi, Divyakant Agrawal,
Ross Wilkinson, Justin Zobel, Michael Fuller,
Giovanna Guerrini, Isabella Merlo, Elisa Bertino,
Robert Rohling, Laurence H. Berman, Andrew H. Gee,
Matthias Felleisen, Shriram Krishnamurthi, Matthew Flatt,
K. Boryczko, Jacek Kitowski, Jacek Moscinski,
Eric G. Wagner, Jesse B. Wright, James W. Thatcher,

Frequent itemsets of length  4 are:
Yervant Zorian, J. M. Portal, Michel Renovell, Joan Figueras,
Mahmut T. Kandemir, J. Ramanujam, Prithviraj Banerjee, Alok N. Choudhary,
Running Time: 5.418997049331665 seconds
```

TASK 2

For the support value of 0.015% many itemsets of length 2 were generated, so I am not displaying them here, you can find them in my output.txt file.

Sample Output for ECLAT algorithm

```
Frequent itemsets of length  3 are:
Mahmut T. Kandemir, J. Ramanujam, Prithviraj Banerjee,
Diego Calvanese, Maurizio Lenzerini, Giuseppe De Giacomo,
Martin Ester, Hans-Peter Kriegel, Jörg Sander,
Germán Vidal, María Alpuente, Moreno Falaschi,
Wolfram Burgard, Sebastian Thrun, Dieter Fox,
Martin Ester, Hans-Peter Kriegel, Xiaowei Xu,
Massimo Violante, Matteo Sonza Reorda, Maurizio Rebaudengo,
Yervant Zorian, Michel Renovell, Joan Figueras,
Yervant Zorian, J. M. Portal, Michel Renovell,
Yervant Zorian, J. M. Portal, Joan Figueras,
J. M. Portal, Michel Renovell, Joan Figueras,
Mahmut T. Kandemir, J. Ramanujam, Alok N. Choudhary,
Mahmut T. Kandemir, Prithviraj Banerjee, Alok N. Choudhary,
Prithviraj Banerjee, J. Ramanujam, Alok N. Choudhary,
Keshav Pingali, Vladimir Kotlyar, Paul Stodghill,
Roland Wismüller, Wlodzimierz Funika, Marian Bubak,
Lipyeow Lim, Jeffrey Scott Vitter, Min Wang,
Henri Prade, Didier Dubois, Salem Benferhat,
Joan Martí, Jordi Freixenet, Xavier Cufí,
Hakan Ferhatosmanoglu, Amr El Abbadi, Divyakant Agrawal,
Ross Wilkinson, Justin Zobel, Michael Fuller,
Giovanna Guerrini, Isabella Merlo, Elisa Bertino,
Robert Rohling, Laurence H. Berman, Andrew H. Gee,
Matthias Felleisen, Shriram Krishnamurthi, Matthew Flatt,
K. Boryczko, Jacek Kitowski, Jacek Moscinski,
Eric G. Wagner, Jesse B. Wright, James W. Thatcher,

Frequent itemsets of length  4 are:
Yervant Zorian, J. M. Portal, Michel Renovell, Joan Figueras,
Mahmut T. Kandemir, J. Ramanujam, Prithviraj Banerjee, Alok N. Choudhary,
Running Time: 3.360443353652954 seconds
```

TASK 3

Strong association rules found using confidence threshold of 75%.

```
{'Yervant Zorian', 'J. M. Portal'}  => {'Michel Renovell', 'Joan Figueras'}

{'Yervant Zorian', 'Michel Renovell'}  => {'J. M. Portal', 'Joan Figueras'}

{'Yervant Zorian', 'Joan Figueras'}  => {'J. M. Portal', 'Michel Renovell'}

{'J. M. Portal', 'Michel Renovell'}  => {'Yervant Zorian', 'Joan Figueras'}

{'J. M. Portal', 'Joan Figueras'}  => {'Yervant Zorian', 'Michel Renovell'}

{'Michel Renovell', 'Joan Figueras'}  => {'Yervant Zorian', 'J. M. Portal'}

{'Mahmut T. Kandemir', 'Prithviraj Banerjee'}  => {'J. Ramanujam', 'Alok N. Choudhary'}

{'J. Ramanujam', 'Prithviraj Banerjee'}  => {'Mahmut T. Kandemir', 'Alok N. Choudhary'}

{'J. Ramanujam', 'Alok N. Choudhary'}  => {'Mahmut T. Kandemir', 'Prithviraj Banerjee'}

{'Prithviraj Banerjee', 'Alok N. Choudhary'}  => {'Mahmut T. Kandemir', 'J. Ramanujam'}
```

TASK 4

No all the association rules may not be interesting so I have used cosine value which is a null-invariant measure to find the interestingness of the above associations and I found that most of the above associations are negatively and some non correlated.

If the cosine value is 1 then the elements are non correlated, if greater than 1 then positively strongly correlated and otherwise negatively strongly correlated.

TASK 5

Strongly negatively correlated co-authors:

```
{'Yervant Zorian', 'J. M. Portal'}  => {'Michel Renovell', 'Joan Figueras'}
Cosine value:  0.8660254037844387

{'Yervant Zorian', 'Michel Renovell'}  => {'J. M. Portal', 'Joan Figueras'}
Cosine value:  0.8660254037844387

{'J. M. Portal', 'Joan Figueras'}  => {'Yervant Zorian', 'Michel Renovell'}
Cosine value:  0.8660254037844387

{'Michel Renovell', 'Joan Figueras'}  => {'Yervant Zorian', 'J. M. Portal'}
Cosine value:  0.8660254037844387

{'Mahmut T. Kandemir', 'Prithviraj Banerjee'}  => {'J. Ramanujam', 'Alok N. Choudhary'}
Cosine value:  0.8660254037844387

{'J. Ramanujam', 'Prithviraj Banerjee'}  => {'Mahmut T. Kandemir', 'Alok N. Choudhary'}
Cosine value:  0.7071067811865476

{'J. Ramanujam', 'Alok N. Choudhary'}  => {'Mahmut T. Kandemir', 'Prithviraj Banerjee'}
Cosine value:  0.8660254037844387

{'Prithviraj Banerjee', 'Alok N. Choudhary'}  => {'Mahmut T. Cosine value:
0.6708203932499369
```