Trabalho Prático II
Mineração de Dados

Claiton H. C. Neisse[1], Deivis C. Pereira[1], Frederico H. dos S. Gassen[1], Mariano D. de Freitas[1], Marlon L. da S. Rodrigues

[1]Curso de Ciência da Computação
Universidade Federal de Santa Maria (UFSM) - Santa Maria - RS - Brasil

{chneisse, dcpereira, fhgassen, mdfreitas, mlrodrigues}@inf.ufsm.br

1. Pré-processamento

O pré-processamento foi feito utilizando scripts Shell com RegEx. Inicialmente, foi passada todas as caixas para caixa-baixa, eliminados os espaços entre os delimitadores, substituídos valores de coluna por números, removidas as colunas descritivas mantendo as de identificação, retiradas as primeiras colunas e as linhas com problemas, feito um split dos subtópicos utilizando a linguagem *AWK* e, por fim, os arquivos auxiliares foram deletados.

2. Transformação

A transformação é realizada no início do script em R. Os arquivos são lidos e agrupados em um Data Frame no formato de transações.

3. Mineração

Dentre as abordagens apresentadas na disciplina, foram utilizadas as abordagens de associação e clusterização. Para associação foi utilizado o algoritmo *apriori* e para clusterização, o algoritmo de *kmeans*. Executar o script *trab2.sh* pré-processa os dados no diretório *dataset* executa o script *trab2.r* e armazena os resultados no diretório *resultados*.

4. Análise dos dados

Usando a abordagem de associação, foram desenvolvidos os comandos a seguir que fornecem as regras que determinam os subtópicos que, historicamente, possuem a maior chance de aceitação de papers.

```
rules <- sort(apriori(data = dados2, parameter = list(support =
0.01, confidence = 0.6)), by = "support")
```

```
subarea <- sort(subset(rules, (size(lhs) == 1) & (lhs %pin%
"subtopics=") & (rhs %in% "status=1")), by = "support")
```

As regras são:

```
     lhs               rhs          support    confidence lift      coun
[1]  {subtopics=25} => {status=1} 0.04067797 0.7317073  1.1109298 240
[2]  {subtopics=26} => {status=1} 0.03186441 0.6064516  0.9207577 188
[3]  {subtopics=31} => {status=1} 0.02830508 0.6987448  1.0608837 167
[4]  {subtopics=42} => {status=1} 0.02644068 0.6872247  1.0433931 156
```

Analisando os dados obtidos, os subtópicos mais aceitos são 25 - “Performance management”, 26 - “Security management”, 31 - “Autonomic and self management” e 42 - “Cloud computing”.

Usando clusterização, foi obtida a quantidade de conferências por ano, subtópicos novos, subtópicos sem nenhum paper e subtópicos com mais submissões rejeitadas. Para as três primeiras informações foi aplicado o *kmeans* nas colunas *year* e *status* com 8 clusters, resultando no código:

```
teste <- kmeans(dados[,4,5],8)
```

Para a quantidade de conferências realizadas por ano, foi utilizado a função *table* para exibir os dados clusterizados em função da coluna *conf*:

```
capture.output(teste$size,teste$centers,teste$withinss,table(teste$
cluster, dados$conf),file = "resultados/conf.txt")
```

```
[1]  501  956  600  577  725 1028  411 1102

  [,1]
1 2014
2 2010
3 2013
4 2011
5 2016
6 2012
7 2017
8 2015

[1] 0 0 0 0 0 0 0 0
```

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	0	0	0	0	0	0	501	0	0	0	0	0	0
2	601	355	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	600	0	0	0	0	0	0	0
4	0	0	327	250	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	485	240	0
6	0	0	0	0	647	381	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	411
8	0	0	0	0	0	0	0	0	704	318	80	0	0	0

Observando os dados, obtemos que o número de conferências por ano não é constante. Entre 2010 e 2012 e em 2016, foram duas conferências por ano. Já em 2013, 2014 e 2017, apenas uma. E no ano de 2015 foram 3 conferências.

Para subtópicos novos e subtópicos sem nenhum paper, foi utilizada a função *table* para exibir os dados clusterizados em função da coluna *subtopics*:

```
capture.output(teste$size,teste$centers,teste$withinss,table(teste$
cluster, dados$subtopics),file="resultados/sub.txt")
```

```
[1] 0 0 0 0 0 0 0 0

    1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
1   7 30 13  2  3  9  4 12 19  3  1  4  3  8  1 21  1  1  5  1  7  9 17  3 42
2  21 53 60  5  4 23 13 14  2  8  2 35 17 15  8 10 15  2  7  2 41 44 32  4 62
3   5 26  4  3  4  5  5 14 13 18  2  9  1 13  0 26 41  5 14  1 10 26  9  3 24
4  10 36 27  2  5  6  4  6  3  6  1 18 19 13  3  8  7  3  1  1 19 18 24 10 38
5   6 58  6  0  4  5  4 37 67 16  6 15  9 10  0 15 22  3  7  1  7 20  9  5 31
6  19 59 20  8  8 24  9 16 11 44  2 25  7 33  2 36 18  2 17  2 20 35 37  6 66
7   2 24  3  1  1  5  5 20 39  5  2  6  0  3  0 10  4  3  4  1  3 12  0  0 14
8   5 48 15  5  5 12 10 37 71 14  3 28  3  6  0 38 40 13 12  1  9 31 41  1 51

    26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
1  23 13  7  3 14 21 15  1  2  4 15  8  4  4  1 13 21  1  0  1  3 14  2  9  1
2  55 14 11  5 44 56 32 15  0 12 27 10  3 25  1 38 21  1  5  6  7 10  4  6  0
3  35  6  1  4 12 19  8  7  2 14 26  5  1  4  0 14 51  3  1  4  0  2  1 14  1
4  28 18  7  5 24 29 19  5  1 13  9  9  2  8  3 28 25  0  2  1  1  2  3  8  1
5  66 19  5  3 19 23  6  8  1 21 16  0  2  2  0 10 29 11  5  0  2 19  1 17  0
6  44 22  5 15 36 57 24  2  1 23 14 22  3 19  3 31 29  3  0  5  8 29  7 15  3
7  17 12  2  1  5  9  5  6  1  6  7  0  0  4  1  1 14  3  0  0  0 12  2  6  0
8  42  8  6 11 15 25  8  4  4 11 31  4  8  8  0 25 37 10  3  6  3 39  7 24  0

    51 52 53 54 55 56 63 64 66 67 68
1   9 23 20  5 15  3  0  0  0  0  0
2   6 13  9  2 16  3  0  0  0  0  0
3   4 20 21  1 23 15  0  0  0  0  0
4   1 10  5  5 13  4  0  0  0  0  0
5   5 22 17  1 19 13  0  0  0  0  0
6  12 20 21  6 20  3  0  0  0  0  0
7   3 23 17  1 11  5  5 10  4 36 15
8  18 65 87 33 37 24  0  0  0  0  0
```

Analisando os dados, foi observado que os subtópicos 63 - “IoT Services”, 64 - “Security Services”, 66 - “Economic Aspects”, 67 - “Software-Defined Networking”, 68 - “Network Function Virtualization” aparecem apenas em 2017 e o subtópico 65 - “Regulatory Perspective” não aparece em nenhuma conferência.

Para a última informação foi aplicado o *kmeans* nas colunas *status* e *topic* com 8 clusters, resultando no código:

```
teste <- kmeans(dados[,5,6],2)
```

Para subtópicos com mais submissões rejeitadas, foi utilizada a função *table* para exibir os dados clusterizados em função da coluna *subtopics*:

```
capture.output(teste$size,teste$centers,teste$withinss,table(teste$
cluster, dados$subtopics),file="resultados/mais.txt")
```

```
[1] 3886 2014
  [,1]
1    1
2    2
[1] 0 0

      1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24
  1  43 194  99  18  20  42  35 109 147  88  10  95  37  84   7 120 113  21  50   4  70 122 111  18
  2  32 140  49   8  14  47  19  47  78  26   9  45  22  17   7  44  35  11  17   6  46  73  58  14

     25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48
  1 240 188  71  26  34 107 167  73  34   8  66  85  36  10  46   6  98 156  17  12  12  13  84  16
  2  88 122  41  18  13  62  72  44  14   4  38  60  22  13  28   3  62  71  15   4  11  11  43  11

     49  50  51  52  53  54  55  56  63  64  66  67  68
  1  63   4  44 134 137  36 109  52   3   6   2  24  10
  2  36   2  14  62  60  18  45  18   2   4   2  12   5
```

Com estes dados, conclui-se que os subtópicos 6 - “Sensor Networks”, 20 - “Legal & ethical issues” e 38 - “Mobile agents” possuem uma taxa de rejeição maior que suas taxas de aceitação, enquanto que, em todos os outros, a relação é inversa (aceitação maior que rejeição).