INF-611 - Juntando Dados - Relatório da Atividade à Distância

Paulo Sigrist / Rafael Sangalli

3 de outubro de 2015

Introdução

Este relatório descreve os resultados obtidos para a atividade a distância da disciplina “INF-611 - Juntando Dados”, bem como os procedimentos executados para chegar a tais resultados.

O procedimento completo executado pode ser resumidamente descrito pelos passos a seguir:

- Obtenção dos dados a partir de uma conexão com o servidor
- Pré-processamento dos dados
    - Conversão dos dados em séries de temperaturas para cada um dos dias contidos nos dados
    - Remoção de valores inválidos (N/A)
    - Remoção de meses com menos de 20 dias de dados
- Cálculo da precisão `P@30` para a busca de uma série de temperaturas utilizando diferentes medidas de distância, em um cenário de busca em que séries de temperatura de um mesmo mês são consideradas relevantes entre si
    - Foram utilizadas as distâncias: L1, L2, L infinito, Canberra e Cosseno.
- Comparação dos resultados obtidos para cada uma das distâncias

Os capítulos a seguir descrevem em mais detalhes cada um dos passos mencionadas acima.

Obtenção dos dados

Conforme o enunciado da atividade, os dados utilizados para na atividade foram obtidos da fonte http://www.ic.unicamp.br/~zanoni/cepagri/cepagri.csv (http://www.ic.unicamp.br/~zanoni/cepagri/cepagri.csv).

Os arquivos foram obtidos a partir de uma conexão direta com o servidor, e carregados em um data frame fazendo uma leitura no formato CSV:

```
## function ()
## {
##     con <- url("http://www.ic.unicamp.br/~zanoni/cepagri/cepagri.csv")
##     cpa <- read.csv(con, header = FALSE, sep = ";", col.names = c("horario",
##         "temperatura", "vento", "umidade", "sensacao"), as.is = TRUE)
## }
```

Veja abaixo um trecho das medidas obtidas:

```
##            horario temperatura vento umidade sensacao
## 1 02/03/2014-19:08        23.7  59.3    77.1     22.6
## 2 02/03/2014-19:18        23.4  59.1    77.9     22.3
## 3 02/03/2014-19:28        23.2  56.7    78.9     22.1
## 4 02/03/2014-19:38        23.0  55.4    79.2     21.9
## 5 02/03/2014-19:48        22.8  52.6    79.7     21.7
## 6 02/03/2014-19:58        22.6  62.6    80.7     21.5
```

Note que os dados foram propositalmente carregados como `character` antes de serem pré-processados de acordo com o próximo passo.

Pré-processamento

No pré-processamento, as medidas de temperatura foram transformadas em séries, onde cada série é um conjunto de 24 números, representando as 24 horas de um dia.

Como é possível observar nos dados de entrada, há mais de uma medida por hora. Para transformar cada dia em uma séria de 24 pontos, foi extraída a média de cada hora. Portanto, o resultado da primeira parte do pré-processamento foi uma matriz de 24 colunas por N linhas, onde N é a quantidade total de dias encontrados nos dados de entrada. Para cada linha da matriz, foi dado um `rowname` que é a própria data a qual a linha se refere.

Veja um exemplo de algumas linhas da matriz:

```
##              [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
## 2014-03-02    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 2014-03-03 20.97 20.62 20.12 19.82 19.62 19.57 19.87 21.33 24.03 25.87
## 2014-03-04 21.67 21.48 21.02 20.75 20.48 20.10 19.88 21.47 24.30 26.57
##             [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20]
## 2014-03-02    NA    NA    NA    NA    NA    NA    NA    NA    NA 23.12
## 2014-03-03 27.77 29.17 30.06 30.48 30.17 29.47 26.65 25.82 24.95 24.93
## 2014-03-04 27.93 29.08 29.87 30.42 30.30 30.38 31.28 30.13 27.65 23.90
##             [,21] [,22] [,23] [,24]
## 2014-03-02 21.98 21.40 21.12 21.03
## 2014-03-03 23.87 23.03 22.47 22.03
## 2014-03-04 24.03 24.03 23.50 22.98
```

Após gerar a matriz, o segundo passo do pré-processamento foi a eliminar valores inválidos, ou seja, aqueles com valor `NA` . Para isso, foi feito o seguinte procedimento:

1. Caso o primeiro dia comece com valor `NA` , este dia é eliminado das séries.
2. Caso o último dia termine com valor `NA` , este dia é eliminado das séries.
3. Para o restante dos dias, cada valor `NA` encontrado é substituído pela média dos pontos imediatamente antes e imediatamente depois.
4. Caso sejam encontrados 2 ou mais valores `NA` em sequência, estes valores são gerados a partir de uma interpolação linear dos pontos imediatamente antes e imediatamente depois.

Por fim, os meses com menos de 20 dias de dados foram removidos do cálculo, pois poderiam influenciar demais no cálculo de precisão.

Terminado este passo, a matriz de 24 colunas por N linhas é retornada como resultado do pré-processamento.

Busca de séries e cálculo da precisão

Foi implementada uma função que, dada uma séria de temperaturas de um dia, eram retornados os resultados que mais se assemelhavam à série de acordo com uma função de distância escolhida. A função recebia como argumento:

- criterio: série de temperaturas de um dia utilizada como critério de buscas
- dias: matriz de 24 colunas e N linhas, onde as colunas são as horas do dia e as linhas são os dias
- funcaoDistancia: função de distância a ser utilizada. Recebe dois vetores como parâmetros
- funcaoOrdenacao: função de ordenação de resultados a ser utilizado. Recebe um vetor como parâmetro.
- numeroRetornados: Quantidade de registros retornados

A função calcula a distância entre o critério e cada um dos dias da matriz, em seguida ordenando os resultados com os mais próximos no topo do resultado. Então são retornados os M primeiros resultados, onde M é o parâmetro `numeroRetornados`. Neste exercício, o parâmetro `numeroRetornados` recebeu sempre o valor 30. Veja abaixo a função que realiza a busca da série:

```
## function (criterio, dias, funcaoDistancia, funcaoOrdenacao, numeroRetornados =
30)
## {
##     distancias <- apply(dias, 1, funcaoDistancia, criterio)
##     distancias <- funcaoOrdenacao(distancias)
##     return(distancias[1:numeroRetornados])
## }
```

Uma vez retornado o resultado, é possível calcular a precisão dos 30 valores retornados (`P@30`). Assim como pede o enunciado da atividade, o critério de acerto é que séries de temperatura de dias de um mesmo mês são consideradas relevantes entre si. A função abaixo foi utilizada para calcular a precisão:

```
## function (anoMes, resultado, dias)
## {
##     anoMesResultado <- obtemAnoMes(names(resultado))
##     anoMesColecao <- obtemAnoMes(rownames(temperaturas))
##     total <- min(sum(anoMes == anoMesColecao), length(resultado))
##     acertos <- sum(anoMes == anoMesResultado)
##     precisao <- acertos/total
## }
```

A função obtemMesAno apenas extrai da data o ano e mês no formato YYYY-mm.

Cálculo das precisões médias

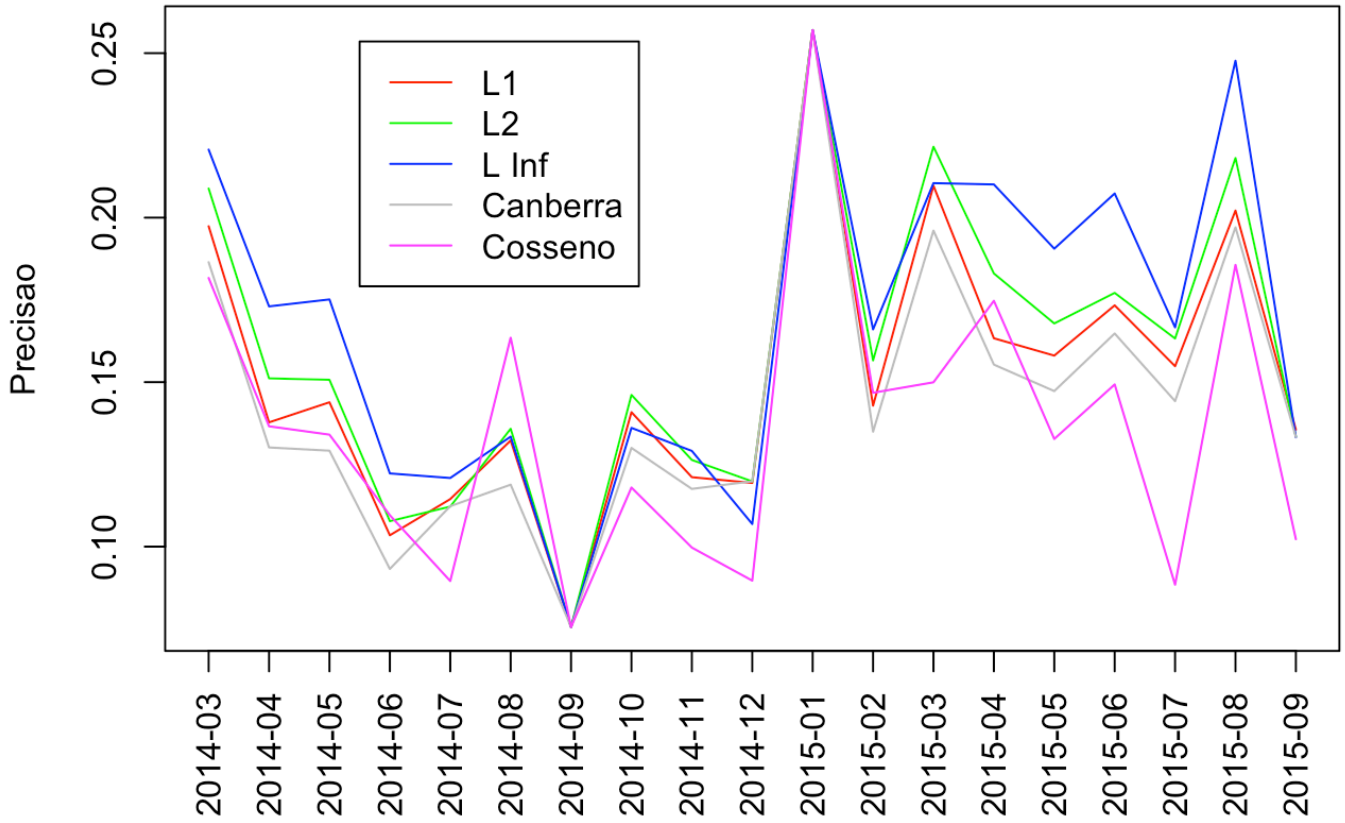
O próximo passo foi calcular a precisão média da coleção, ou seja, foram feitas buscas e cálculos de precisão utilizando como critério cada uma das séries da coleção, e em seguida extraída a média dessas precisões. Foi feito tanto a média de precisão `P@30` por mês como a média de precisão `P@30` geral. O cálculo de precisão média foi feito com diferentes funções de distância, e a comparação entre os resultados obtidos para cada distância pode ser conferida no próximo capítulo.

Comparação dos resultados obtidos para cada uma das distâncias

Para a realização das buscas foram utilizadas as seguintes funções para calcular a distância:

- L1
- L2
- L Infinito
- Distância de Canberra
- Similaridade por Cosseno

O gráfico abaixo compara a precisão média `P@30` por mês obtida para cada uma das funções de distância utilizada:



Como é possível notar no gráfico, nenhuma das funções de distância utilizada se destacou muito em relação às outras, e a precisão máxima obtida não passou de 0.25. Ou seja, independente da função de distância utilizada, a precisão do resultado apresentado é abaixo de 25%, mostrando a dificuldade de encontrar as temperaturas de um determinado mês dada a temperatura de um dia do mês.

Um outro dado interessante é que o mês de janeiro/2015 foi o que apresentou a melhor precisão. Ao analisar os dados de temperatura, é possível notar que a média de temperatura do mês de janeiro é bem mais elevada do que outros meses, e portanto, o que tornou possível obter uma precisão melhor para este mês. Veja abaixo que janeiro é o único mês na faixa de 25 graus, enquantos para outros meses sempre há pelo menos dois meses com temperaturas próximas entre si:

```
##         mediaMensal
## 2014-03    23.78719
## 2014-04    21.95708
## 2014-05    19.52855
## 2014-06    19.54123
## 2014-07    18.81823
## 2014-08    20.09290
## 2014-09    22.66704
## 2014-10    23.85877
## 2014-11    23.19394
## 2014-12    23.64341
## 2015-01    25.51657
## 2015-02    23.50825
## 2015-03    22.28965
## 2015-04    21.89768
## 2015-05    19.11822
## 2015-06    18.46698
## 2015-07    18.52324
## 2015-08    20.60309
## 2015-09    22.14896
## 2015-10    23.93711
```

Por fim, podemos também comparar a precisão média geral `P@30` para cada uma das distâncias, independente do mês avaliado:

```
##        L1        L2      LInf  Canberra   Cosseno
## 0.1517300 0.1536209 0.1555092 0.1528360 0.1372934
```

Note que L infinito foi um pouco mais eficaz do que as outras medidas de distância, o que talvez seja um indício de que é mais eficaz comparar temperaturas diárias utilizando apenas a temperatura mais alta ao invés de levar em consideração todas as temperaturas do dia.

Dificuldades enfrentadas

Foram duas as principais dificuldades enfrentadas:

- A realização do pré-processamento dos dados é uma parte do trabalho que necessita de uma análise intensa e um esforço considerável. Isso ocorre porque é necessário conhecer os dados para fazer o pré-processamento, assim sendo possível eliminar dados que possam prejudicar o resultado e determinar um método para adicionar valores que não tenham sido informados, mas que são importantes para o processamento. Além disso, há diversos casos diferentes a serem tratados que são totalmente dependentes de situações específicas que talvez só sejam notadas com o recebimento de novos dados. Um exemplo em relação a isso: grande parte deste relatório foi feita durante o final do mês de setembro, quando todos os meses analisados tinham mais de 25 dias de dados. Ao finalizar o relatório, em outubro, já havia alguns poucos dias do novo mês, e então ao executar o algoritmo novamente, a precisão `P@30` de outubro era consideravelmente maior que os outros dias. Isso se deu ao fato de que outubro tinha apenas três dias para serem avaliados. Para evitar esse problema, foi adicionada uma nova etapa que retirava dos dados os meses com menos de 20 dias.
- Independente da função de distância utilizada, nenhuma das precisões `P@30` obtida foi

consideravelmente boa, sendo que isso não era conhecido antes de iniciarmos a análise. Isto demostra que nem sempre é trivial, ou mesmo possível, chegar a uma boa eficácia dos resultados para certas situações.

Conclusão

Pode-se concluir que ao iniciar um trabalho na área de recuperação de informação, é bastante difícil saber no início se será possível chegar a um resultado satisfatório.

Neste caso específico das temperaturas, é possível pensar intuitivamente que um determinado mês possui temperaturas bem semelhantes a outros meses próximos, e essa proximidade da temperatura se torna ainda maior em países como o Brasil em que a temperatura não tem variações extremas ao longo do ano. Estes fatores são um indício que, de fato, a recuperação de temperaturas de um mesmo mês não é um trabalho tão simples.

Talvez com uma análise mais profunda e a utilização de outras técnicas fosse possível chegar em uma eficácia maior. Um exemplo do que poderia ser feito é utilizar outras características da temperatura diária no modelo vetorial, tal como a média, mediana, máxima e mínima do dia, dentre outros valores. Ou ainda, se aceitável, substituir o critério de relevância: ao invés de considerar temperaturas de um mesmo mês como relevantes entre si, considerar todas as temperaturas de uma mesma estação do ano Poderíamos também usar outras variáveis do Cepagri na criação do modelo como, por exemplo, as medidas de vento e sensação térmica. Mesmo assim, isto não garante que os resultados obtidos seriam bons.