Trabalho Prático 1
Visualização de Padrões em Comunidades

Larissa Leijôto, Péricles Alves, Rubens Emilio, Vinícius Garcia

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

[leijoto,periclesrafael,rubens,garcia]@dcc.ufmg.br

1. Introdução

Visualizar informação não é uma tarefa simples: além dos problemas inerentes ao processamento dos dados, construir visualizações claras e objetivas requer conhecimentos básicos sobre o processo cognitivo. Deixando de lado as etapas de coleta e tratamento da base de dados – que são, em si, problemas complexos –, existem diversas estratégias para extrair informações relevantes de conjuntos de dados.

Diversos algoritmos de áreas como mineração de dados, aprendizado de máquina e computação natural, processam os dados a fim de identificar padrões implícitos. O volume dos dados e a existência de ruídos são exemplos de problemas recorrentes nestas áreas de pesquisa. Ao final do processamento, é necessário exibir as informações de forma objetiva, valorizando os resultados obtidos, e permitindo novas conclusões sobre os dados.

A tarefa de visualizar dados, além de lidar com os desafios discutidos acima, ainda depende de conhecimentos acerca do processo cognitivo. É preciso ter um mínimo de noção sobre como o ser humano percebe e interpreta informações visuaos. Só então é possível decidir como exibir resultados pré-computados, ou ainda, como exibir os dados iniciais para que o leitor consiga, com relativamente pouco esforço, tirar conclusões valiosas.

Neste trabalho, será implementada uma técnica de visualização de padrões em grafos. A visualização será feita sobre conjuntos disjuntos de pessoas (comunidades), que compartilham um conjunto predefinido de lugares que costumam visitar. O resultado do trabalho é uma representação gráfica para cada conjunto de pessoas. Assim, espera-se que, ao analisar mais de uma comunidade, seja possível identificar padrões visuais entre as mesmas.

2. Padrões em Comunidades

Os dados a serem analisados contêm comunidades compostas por pessoas, que, por sua vez, podem visitar diferentes lugares. O objetivo principal de representar tais comunidades visualmente é conseguir identificar, com pouco esforço, correlações entre pessoas de comunidades distintas.

2.1. Donut Graphs

Para representar cada indivíduo, será utilizada uma técnica de representação do tipo gráfico de pizza, denominada *donut graph*. Este tipo de gráfico apresenta, de forma bastante clara, a frequência com que determinado indivíduo visita diferentes lugares.

Além da espessura de cada fatia do gráfico indicar a frequência de visitas, é possível associar cores a cada lugar. Isto possibilita a identificação de padrões visuais, principalmente quando diferentes gráficos (indivíduos) são comparados.

2.2. Posicionamento e Arestas

O posicionamento de cada pessoa em sua respectiva comunidade é utilizado para representar a semelhança entre pessoas de diferentes comunidades. Desta forma, se uma pessoa tem hábitos similares aos de uma pessoa de outra comunidade, ou seja, se visitam os mesmos tipos de lugares, então existe uma correlação entre elas, e o posicionamento das mesmas no espaço será igual.

Além disso, sempre que pessoas de uma mesma comunidade visitam um mesmo lugar, é desenhada uma aresta entre as mesmas, facilitando a identificação de mais padrões visuais. Neste caso, os padrões serão referentes às conexões entre cada *donut graph*.

2.3. * Heatmap

A fim de explorar outras técnicas de visualização, procuramos alguma abordagem que sumarizasse as correlações entre as pessoas de diferentes comunidades. Com isto, decidimos implementar o modelo de visualização por *heatmap*, muito úteis na representação de grandes quantidades de dados, e sem o problema de sobrepor informações.

No caso da análise de similaridade entre comunidades, criamos uma representação para cada indivíduo. Uma pessoa é descrita como um vetor da frequência com que a mesma visita lugares de tipos diferentes. Aos valores de frequência são associadas diferentes gradações de cores.

3. Desafios e Detalhes de Implementação

Após a implementação das visualizações, foi possível tirar algumas conclusões a respeito dos dados analisados. Tais conclusões são obtidas a partir de padrões visuais nas imagens criadas.

3.1. Posicionamento de indivíduos nos gráficos de tipo Donut

Ao plotar os plotar os gráficos de tipo *donut*, foi preciso definir posições padrão para cada indivíduo. A fim de minimizar a poluição visual, decidimos posicionar os indivíduos dispondo-os sobre um círculo. Desta forma, tanto conseguimos evitar sobreposições – um problema inerente à técnica de visualização utilizada –, como conseguimos melhorar o aspecto visual das arestas nos gráficos.

3.2. Detecção de outliers

Os gráficos gerados representam comunidades de indivíduos. O posicionamento de pessoas de comunidades diferentes, em seus respectivos gráficos, depende diretamente da similaridade entre tais pessoas. No entando, existem pessoas que não apresentam padrões de visitas similares aos de nenhuma das outras pessoas.

O gráfico abaixo mostra a soma das distâncias de cada indivíduo em relação a todos os outros. A partir do gráfico, é possível perceber que alguns indivíduos possuem distâncias muito altas em relação aos outros, e, consequentemente, têm hábitos não similares aos de qualquer outra pessoa.

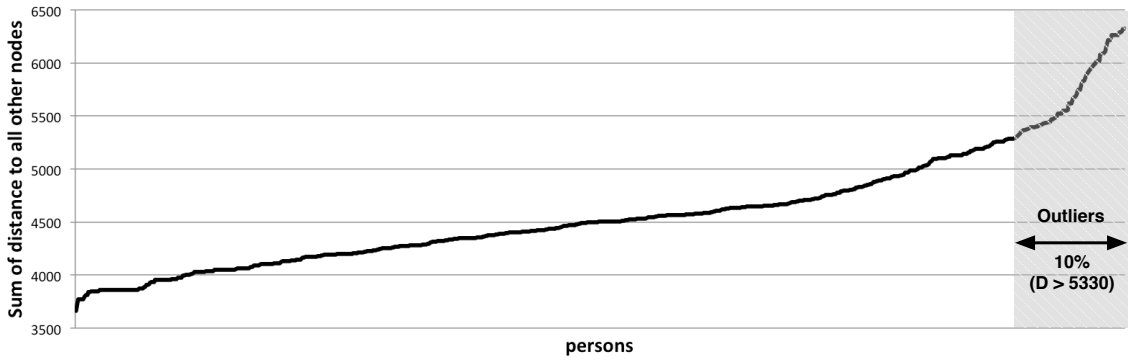

Figura 1. Soma de distâncias de todas para todas as pessoas

3.3. Filtragem de arestas

Sobre o resultado inicial dos gráficos, notamos que os grafos de *donuts* eram completos, ou seja, tinham arestas entre todos os indivíduos. A fim de reduzir ainda mais a poluição visual, decidimos parametrizar a inclusão de arestas nos gráficos. Assim, para que exista uma aresta entre dois indivíduos, os mesmos têm que ter visitado determinado lugar uma quantidade mínima de vezes – definida na interface web desenvolvida.

3.4. Heatmap

Após a implementação da visualização por *heatmap*, podemos perceber que as pessoas com mesmo identificador eram, no geral, agrupadas como similares. Além disso, vimos também que esta abordagem produz resultados visualmente mais claros e intuitivos que a original – *donut graphs*.

3.5. Métricas de similaridade

Para o trabalho, foram implementadas duas métricas de similaridade. A primeira, sugerida na especificação, foi a distância euclideana. A segunda métrica utilizada foi a similaridade de cosseno. Em relação aos resultados, ambas as métricas produziram resultados visuais similares, apesar da métrica de similaridade de cosseno ser mais robusta sobre dados com alta dimensionalidade.

4. Conclusão

Durante a análise visual de dados, o usuário é o responsável por identificar, distinguir, categorizar e comparar informações a ele apresentadas. Desta forma, uma boa estratégia de visualização é aquela que facilita o trablho do usuário, ou seja, que propicia a identificação de padrões com pouco esforço.

Com isso, notamos durante o trabalho, que a técnica de *Heatmap*, relativamente mais simples que os *Donut graphs*, tem resultado visual mais representativo. Ou seja, é possível identificar, de forma mais intuitiva, os padrões existentes entre as comunidades.

Foram implementadas duas interfaces web, uma para cada visualização desenvolvida. Para acessar o trabalho principal, feito com os gráficos de *donuts*, o link utilizado é `http://dcc.ufmg.br/~periclesrafael/tp-visu/visualization.html`. E, para o trabalho extra, o mesmo foi hospedado em `http://dcc.ufmg.br/~rubens/dv/heatmap.html`.