Instituto de Ciências Matemáticas e de Computação USP-ICMC

Tópicos Avançados em Inteligência Artificial - SCC0532

Prof. Dr. Alneu de Andrade Lopes

Predição de Links

Gabriela de Oliveira 7572867
Vinicius Alvarenga Lovato 7696455

Introdução

Atualmente, diversos problemas podem ser representados através de redes; sistemas de transporte aéreo, linhas férreas, problemas de logística, e também as redes sociais são alguns exemplos onde esta representação é muito utilizada. As redes são representadas por grafos que contém como elementos principais vértices e arestas. As arestas representam uma relação entre as entidades da rede, no caso das redes sociais como o *facebook* são representações de amizade entre as pessoas.

Um dos problemas importantes na área de redes é a predição de links, que consiste em prever futuras conexões entre os nós existentes de uma rede utilizando os dados já existentes. A predição de links tem sido aplicada em diversas áreas, tais como sugestão de colaboração de cientistas, interações entre proteínas e até mesmo detecção de redes de terrorismo [1].

Os métodos de predição de links apresentados nesta monografia são o Common Neightborhood e Jaccard, ambos serão apresentados com maiores detalhes nas próximas seções

O objetivo deste trabalho é extrair informações de uma sub-rede do *facebook*, representada por um arquivo no formato texto. Além da extração de informações gerais sobre a rede, serão também implementados e avaliados os preditores de links desenvolvidos assim como a influencia da realização de pre-processamento no conjunto de dados no resultado final.

Esta monografia está organizada nas seguintes seções: descrição dos dados, técnicas utilizadas, resultados e conclusão

Descrição dos Dados e Pré-Processamento

O conjunto de dados utilizado neste trabalho é um arquivo de texto que contém conexões entre os nós de uma rede, no caso, uma sub rede do *facebook.* Cada linha do arquivo contém dois números que representam uma aresta, isto é, uma ligação que representa uma amizade entre as pessoas da rede. Esta ligação de amizade no *facebook* é mútua, ou seja, se A esta conectado a B então B esta conectado a A.

Abaixo temos um exemplo de entradas do arquivo utilizado

```
2346 2025
2140 2428
```

As entradas acima criam uma ligação, ou amizade, entre as pessoas de número 2346 e 2025, assim como entre as pessoas 2140 e 2428.

Como a quantidade de arestas e nós do grafo analisado é muito grande, foi escolhida a matriz de adjacência para representar a rede. A linguagem utilizada para o desenvolvimento da aplicação foi Javascript, pois a linguagem encapsula diversas funções facilitando a manipulação do grafo.

Inicialmente é carregado para a memória o arquivo que contém as conexões da rede, e posteriormente é feito o *parsing*. De forma otimizada, para cada nó que aparece a esquerda da linha é criada uma entrada na lista que contém todos os nós da rede; cada nó presente a direita é então adicionado a lista de adjacência do seu respectivo par.

Como exemplo podemos tomar as seguintes entradas de um arquivo:

```
1 2
2 1
3 1
1 3
```

As entradas acima geram a seguinte lista de adjacência:

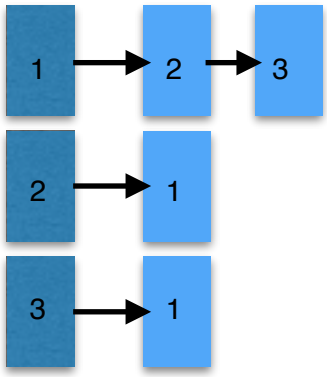

Figura 1

Antes de efetuar o pre-processamento dos dados, foram desenvolvidos algoritmos para extrair algumas métricas da rede, tais como grau de um determinado vertice, média de grau da rede e coeficiente de agrupamento local e global. Os algoritmos e métodos serão explicados com maiores detalhes posteriormente

Métricas	Valores
Número de Nós da Rede	748
Número de Conexões	30025
Média de grau da rede	80.2820
Coeficiente Agrupamento Médio	0.63540

Tabela 1

Grau dos Vértice

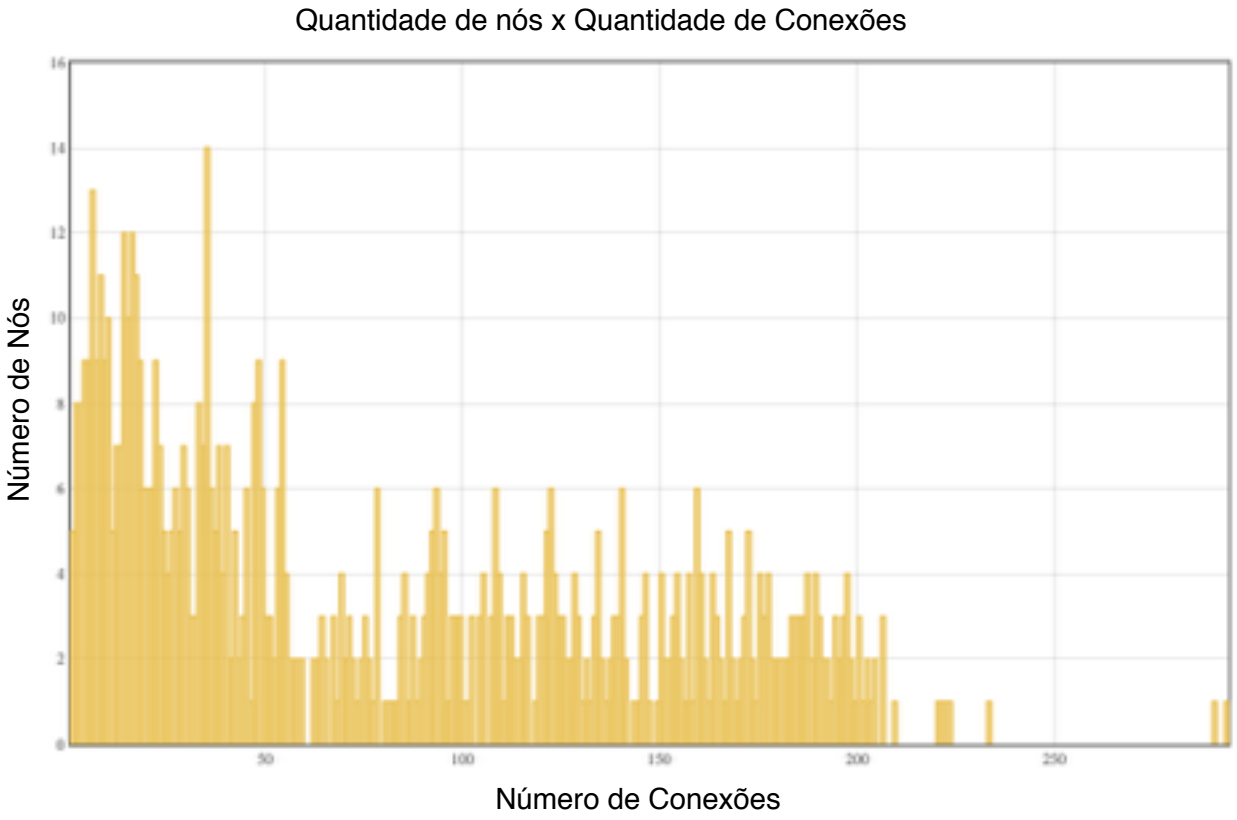

Gráfico 1

Através do gráfico acima percebemos que existe uma grande quantidade de nós com menos de 50 conexões, e existem poucos nós com numero de conexões acima de 200. Estes nós com muitas conexões são denominados hubs.

Hubs são considerados nós importantes em uma rede complexa,; estes nós podem influenciar o posicionamento e conexões de novos integrantes da rede, além de funcionar como ponto de dissipação de informações, no caso de redes reais estes nós possuem a semântica de grandes aeroportos, pontos de disseminação de doenças, termos de procura na web, entre outros.

No caso estudado os hubs são as pessoas que possuem a maior quantidade de amigos na rede social, e funcionam como ágeis dissipadores de informações, pois estão conectados com boa parte da sub-rede analisada.

Através do gráfico percebemos que os hubs estão muito distantes da média de grau da rede, que equivale a cerca de 80 conexões, o que pode gerar ruídos nos resultados dos algoritmos de predição de link. Podemos então efetuar a remoção desses nós da rede em uma etapa de pre-processamento de modo a alterar as métricas para obter outros resultados ao utilizar os algoritmos de predição.

Técnicas Utilizadas

Propriedades da Rede

Grau do Vértice:

O grau de cada vértice é calculado a partir do tamanho da lista conectada a cada nó da lista de adjacência, portanto no exemplo da *figura 1,* temos que o grau de cada vértice é dado por:

`Grau do Vértice 1:` $g = |(2,3)| = 2$

`Grau do Vértice 2:` $g = |(1)| = 1$

`Grau do Vértice 3:` $g = |(1)| = 1$

A média de grau é feita através da média dos valores de grau de vértice de cada nó do grafo, utilizando o mesmo exemplo da figura 1 temos que a Média do Grau é de:

$$Md = \frac{(2+1+1)}{3} = \frac{4}{3}$$

Coeficiente Agrupamento Local:

O coeficiente de agrupamento local de um nó da rede mede o quão próximo seus vizinhos estão, medindo o quão densa são as ligações próximas ao nó avaliado. O valor deste coeficiente para um determinado nó pode ser dado pelo numero de conexões que seus vizinhos fazem divido pelo numero de conexões que poderiam existir. Observemos o seguinte exemplo

Dado um grafo com 4 nós conectados da seguinte maneira.

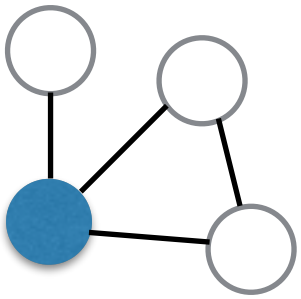

Figura 2

Devemos então verificar quantas são as possíveis conexões no qual os vizinhos do nó azul analisado poderiam ter, no caso temos as seguintes possíveis conexões marcadas em vermelho:

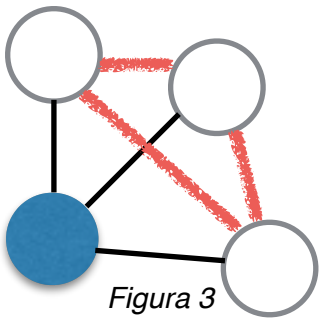

Figura 3

Dessa forma podemos notar que o número total de conexões possíveis que poderia existir entre os vizinhos do nó azul é de T = 3. Dado que o numero de conexões existentes é de 1 temos que o coeficiente de agrupamento local do nó azul é de C = 1/3.

Umas das maneiras de verificar as conexões é por meio da contagem de triângulos, porem a maneira mais simples encontrada para desenvolver o algoritmo foi utilizar análise combinatória. Podemos chegar ao resultado do total de possíveis combinações da utilizando os seguintes passos:

- Queremos todas as possíveis combinações de conexão entre os nós vizinhos ao vértice azul
- A ordem das conexões não é importante

Dado as duas afirmações acima temos então que o total de possíveis conexões, ou triângulos, é dado por:

$$\mathrm{C}(n, s) = \left(\begin{matrix} n! \\ s! \cdot (n - s)! \end{matrix} \right)$$

Aplicando a fórmula de combinatória ao problema do coeficiente de agrupamento queremos então combinar n vértices 2 a 2, no caso do exemplo n = 3, portanto:

$$\mathrm{C}(3, 2) = \left(\begin{matrix} 3! \\ 2! \cdot (3 - 1)! \end{matrix} \right) = 3$$

Coeficiente Agrupamento Médio:

O coeficiente de agrupamento médio foi calculado utilizando a seguinte formula,

$$\mathrm{C} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{C}_i$$

no qual C_i é o coeficiente de agrupamento local de cada vertice do grafo analisado

Common Neighbours e Jaccard

Neste trabalho foram implementados dois algoritmos de predição de links, um deles utilizando a medida de predição de link Common Neighbors e outro utilizando a medida Jaccard. Estes métodos são baseados no cálculo de propriedades dos vizinhos de um determinado nó que esta sendo avaliado. Podemos denotar por exemplo, o conjunto de vizinho do nó x como r(x) e o conjunto de vizinhos do no y como r(y). Os métodos baseados em vizinhança levam em consideração o tamanho da sobreposição entre os conjuntos r(x) e r(y) para avaliar o quão próximos os nós estão. Desta forma

se esta sobreposição for suficientemente grande, então podemos dizer que existe uma probabilidade alta desse dois nós estarem conectados em estados futuros da rede.

A medida Common Neighbours verifica a correlação entre os vizinhos de um dado nó x e y em um determinado estado da rede. Esta pontuação é dada pelo por:

$$y = | \tau(x) \cap \tau(y) |$$

O coeficiente de Jaccard é uma métrica utilizada para medir similaridade e diversidade entre conjuntos, é medida a probabilidade do nó x e y possuir uma determinada característica C, onde no caso a característica são simplesmente os vizinhos do nó. Temos então que a pontuação é dada por:

$$y = \frac{| \tau(x) \cap \tau(y) |}{|\tau(x) \cup \tau(y)|}$$

Para gerar os rankings utilizando as medidas Common Neighbours e Jaccard são gerados os conjuntos de nós vizinhos de cada nó da lista de adjacências. Posteriormente combinamos cada conjunto gerado com o conjunto de todos os outros nós de forma a criar combinações de possíveis conexões futuras, conexões estas que ainda não estão presentes no estado atual da rede. Podemos demonstrar como os rankings são gerados por meio do seguinte pseudocódigo:

```
Para cada conjunto r da lista de adjacência
      Para cada conjunto s da lista de adjacência diferente de r
            Se a combinação de r e s gera um link ainda não existente
                  calcula CN de r x s
                  calcula Jaccard r x s
Ordena lista CN
Ordena lista Jaccard
```

Após gerar as listas utilizando CN e Jaccard, posteriormente criamos o ranking através do ordenamento destas listas, dessa forma os links com maior probabilidade de se formar são posicionados em seu topo. Como o ranking gerado pode se tornar muito grande obtemos somente os 10% com maior pontuação. Vale lembrar que os rankings

são gerados utilizando 90% do conjunto de dados, e os outros 10% são utilizados para verificar o quão preciso foram os resultados dos preditores. Mais detalhes sobre o processo de treinamento e teste serão demonstrados na parte de resultados

Durante o cálculo das listas de predições algumas outras estatísticas também são geradas; o grau de cada nó pode ser calculado obtendo o tamanho da lista de adjacência de cada nó e a media é calculada através da soma de todos estes valores divididos pelo numero de nos que a rede contém.

Resultados

Avaliação dos Algoritmos de Predição

Para avaliar os preditores de links desenvolvidos neste trabalho necessitamos conhecer o estado futuro da rede, porém não é possível obter esta informação. Desta forma o *workaround* normalmente usado é remover alguns links da rede original de forma e tentar preve-los usando os preditores. O tipo de avaliação utilizado foi o *random sub-sampling* que consiste em selecionar de forma aleatória uma parte dos dados para treino e o restante dos dados para teste. O conjunto de treino possui cerca de 90% dos links originais da rede, os 10% restantes são separados e comparados com o ranking de possíveis links gerados pelo Common Neighbour e Jaccard.

O grafo do conjunto de dados utilizado possui um grande número de vertice e arestas, porém podemos observar o topo do ranking gerado pelos algoritmos Common Neighour e Jaccard

#	Common Neighbour	Rede Futura Contem Link	Jaccard	Rede Futura Contem Link
1	[1983,2266] = 179	NAO	[1961,2487] = 1	NAO
2	[2244,2464] = 174	SIM	[2080,2358] = 1	NAO
3	[2123,2324] = 173	SIM	[2244,2464] = 0.8405	SIM
4	[2218,2244] = 173	SIM	[2218,2244] = 0.8277	SIM
5	[2150,2206] = 171	SIM	[2150,2206] = 0.8181	SIM
6	[2206,2324] = 170	NAO	[2078,2593] = 0.8155	NAO
7	[2201,2266] = 169	NAO	[2123,2324] = 0.81220	SIM
8	[2078,2593] = 168	NAO	[2201,2206] = 0.81159	NAO
9	[2201,2206] = 168	NAO	[2590,2607] = 0.8029	SIM
10	[2088,2369] = 167	NAO	[2059,2131] = 0.7990	SIM

Tabela 2

Apesar da quantidade de dados pequena presente na tabela, os valores apresentados são as 10 melhores predições de cada algoritmo, podemos então considerar as informações da tabela como tendo alta relevância. Podemos observar que o Common Neihgbour teve um acerto de apenas 40% enquanto o Jaccard teve um acerto de 60%.

Conclusão

Utilizando predição de links é possível avaliar a probabilidade do surgimento de um relacionamento entre dois nós de uma rede complexa, para isto algumas medidas podem ser utilizadas para verificar a similaridade entre nós e respectivamente a chance de existir uma conexão futura entre eles. Neste trabalho foi demonstrado como extrair algumas métricas da rede, tais como coeficiente de agrupamento, grau de vertice, entre outros. Foi também demonstrado o funcionamento de algumas medidas de predição, tais como Common Neighbour e Jaccard, avaliando o desempenho dos mesmo. Notou-se que com o auxilio de medidas relativamente simples é possível prever o surgimento de links em uma rede de forma satisfatória.

Apesar da simplicidade das medidas utilizadas, o tempo médio de execução e processamento dos algoritmos para extrair informações da rede é alto e tende a crescer muito conforme o tamanho da rede aumenta, deixando claro que esta área exige um poder computacional muito elevado.

Referências

[1] Armada, Marcius, e Revered Kate. “Predição Semântica de Links: algoritmos e aplicações” Universidade Federal do Estado do Rio de Janeiro

[2] Liben−Nowell, David, and Jon Kleinberg. "The link−prediction problem for social networks." *Journal of the American society for information science and technology* 58.7 (2007): 1019-1031.