Resumo Estendido para o 2º Encontro de Analista de Redes Sociais

Uma Medida de Coesão Baseada em Sequências de Coberturas por k-cliques

Luís Cavique
Universidade Aberta, lcavique@univ-ab.pt

Armando B. Mendes
Universidade dos Açores, amendes@uac.pt

Palavras-chave: extracção de conhecimento em grafos, redes sociais

1 Introdução

Depois da comunicação de Tim Berners-Lee na conferência internacional de World Wide Web WWW2006 sobre as três idades da Web, as redes sociais associados à Web 2.0, suscitaram uma explosão de interesse, na tentativa de melhorar a socialização e na criação de novos modelos de gestão dos conhecimentos.

O termo "redes sociais" foi cunhado por Barnes em 1954, no entanto, a visualização através de grafos, chamados sociogramas, foi apresentado por Moreno. Esta área científica da sociologia tenta explicar como emerge a liderança, como se criam alianças e conflitos, como se difunde a inovação/epidemia e como é que a estrutura de um grupo afecta a sua eficácia.

Um importante desenvolvimento sobre a estrutura das redes sociais teve origem numa experiência realizada pelo psicólogo americano Stanley Milgram. A experiência de Milgram consistiu em enviar cartas de pessoas no Nebraska, no Centro-Oeste, para serem recebidas em Boston, na costa leste dos EUA, com a seguinte particularidade - as pessoas eram instruídas a passar as cartas de mão em mão, a alguém que conheciam, até chegarem ao destino. As cartas que chegaram ao destino foram passadas por cerca de seis pessoas. Milgram concluiu que, os americanos não estão a mais de seis passos entre si. Esta experiência deu origem ao conceito de “seis graus de separação” e ao conceito de “pequeno mundo”.

Um exemplo interessante dos “pequenos mundos” é o "Número de Erdos" . Erdos é um dos mais prolíficos matemáticos de todos os tempos, sendo autor de mais de 1500 artigos publicados com mais de 500 co-autores. Assim, se Erdos fôr classificado como o nó zero, os investigadores que trabalharam com ele são Erdos número 1. Os co-autores de Erdos número 1 são conhecidos como Erdos número 2, e assim por diante, construindo um dos mais antigos “pequenos mundos”. O trabalho de Erdos e Renyi apresenta interessantes propriedades dos grafos aleatórios. O interesse foi reavivado recentemente com o modelo de Watts e Strogatz, publicado na revista Nature, que apresenta novas propriedades dos pequenos mundos.

Os analistas de redes sociais precisam de realizar inquéritos de cada pessoa sobre os seus amigos, pedir a sua aprovação para publicar os dados e acompanhar a população durante anos. Por outro lado, as redes sociais como o LinkedIn, Facebook, Hi5 e o MySpace podem fornecer os dados necessários sem este tipo de esforço.

A extracção de conhecimento de grafos pode ser definida como a arte e a ciência de encontrar informação útil e grafos, como padrões e "outliers", fornecidos respectivamente, por dados repetidos ou dados esporádicos existentes em grafos de grandes dimensões ou de redes complexas.

Neste trabalho, propomos criar novas medidas com base na cobertura do grafo com k-cliques, com o objectivo de obter uma visão geral do grafo. Na secção 2, revemos os conceitos de redes sociais, detalhando as estruturas de subgrupos coesos. Na secção 3 apresentamos o algoritmo de duas fases que analisa em primeiro lugar para subgrupos coesos e, em segundo lugar descobre o conjunto mínimo de subgrupos coesos que cobrem todos os vértices. Na secção 4 são apresentados os resultados computacionais e exemplos. Finalmente, na secção 5, são apresentadas as conclusões.

2 O Algoritmo de Duas-Fases

As medidas utilizadas na análise de Redes Complexas e "Graph Mining" são baseadas em procedimentos de baixa complexidade computacional, como o diâmetro do grafo, o grau de distribuição dos nós e a verificação da conectividade, subestimando o conhecimento da estrutura das componentes do grafo. Neste trabalho, propomos uma medida baseada na cobertura do grafo com k-cliques, com vista a compreender melhor a estrutura do grafo, combinando a análise de subgrupos coesos com o conceito de "pequenos mundos".

Com o fim de encontrar a sequência de coberturas mínimas por k-cliques é proposto um algoritmo de duas fases. Na primeira fase, introduzimos o conceito de grafo k-G(V, E). Dado um grafo G(V, E), um grafo k-G(V, E) é a transformação do grafo G(V, E), tal que para cada $i,j \in V$, a distância $d(i,j) \leq k$. Para encontrar todas as k-cliques maximais do grafo, usamos uma simples transformação do grafo e, em seguida, utilizamos um algoritmo multi-partida da clique máxima. Esta fase tem como objectivo gerar todas as k-cliques maximais do grafo. Na segunda fase, cada subconjunto mínimo de k-cliques é escolhido para cobrir todos os vértices do grafo.

Procedimento 1: Algoritmo de duas-fases para encontrar a cobertura por k-cliques
Entrada: distância k e grafo G(V, E)
Saída: cobertura por k-cliques
1. Encontrar todas as k-clique maximais do grafo G:
1.1. Transformar o grafo num k-G(V,E)
1.2. Aplicar o algoritmo multi-partida da clique máxima
2. Aplicar o algoritmo da cobertura de conjuntos com k-cliques

3 Resultados Computacionais

Para a implementação computacional do algoritmo algumas escolhas têm de ser feitas, como o ambiente computacional, as medidas de "graph mining" e as instâncias de teste.

Os programas foram escritos em linguagem C utilizando o compilador Dev-C++. Os resultados computacionais foram obtidos num processador 2.53GHz Intel Core 2Duo com 4,00 GB de memória principal, a funcionar sob o sistema operativo Windows Vista.

A medida de "graph mining" proposta é a sequência de cardinalidades das coberturas por k-cliques de um grafo, que se apresenta no procedimento seguinte. O algoritmo das Duas-Fases (geração de k-cliques e cobertura do grafo) é repetido para valores de k=1 até k=diam(G), com vista a encontrar a sequência de coberturas mínimas de k-G(V, E). Quando k=diam(G), ou a cardinalidade da cobertura por k-cliques é igual à unidade, estamos na presença de um "pequeno mundo". Por isso dizemos que, a sequência de coberturas de k-G(V, E) combina a análise de subgrupos coesos com o conceito de "pequenos mundos".

Nos resultados computacionais do Procedimento 0, geração da sequência de cardinalidades das coberturas por k-cliques foram utilizados dois grupos de conjuntos de grafos, os grafos Erdos e alguns ficheiros do concurso da clique DIMACS.

Tabela 1: Sequência das Cardinalidades das Coberturas por k-cliques

grafo	nr nós	diâmetro	cardinalidade da cobertura por k-cliques									
			k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=9	k=18	k=40
Erdos-97-1	472	6	9	8	7	7	4	1	--	--	--	--
Erdos-98-1	485	7	8	8	7	5	1	1	1	--	--	--
Erdos-99-1	492	7	8	8	7	7	1	1	1	--	--	--
brock200_1	200	2	24	1	--	--	--	--	--	--	--	--
brock200_2	200	2	26	1	--	--	--	--	--	--	--	--
brock400_1	400	2	26	1	--	--	--	--	--	--	--	--
brock400_2	400	2	23	1	--	--	--	--	--	--	--	--
c-fat200-1	200	18	16	11	9	8	7	7	6	5	1	--
c-fat200-2	200	9	9	7	5	4	3	3	3	1	--	--
c-fat500-1	500	40	16	12	9	7	7	6	6	4	3	1

Para a análise de cada grafo, considerámos o número de vértices, o diâmetro e a cardinalidade do conjunto de k-cliques que cobrem todos os nós, variando k de 1 até ao valor do diâmetro. Os grafos Erdos-98-1 e Erdos-99-1, com diâmetro 7, são cobertas apenas com um 5-clique. Estes valores exemplificam a diferença entre o k-clique e o k-clã; estes grafos são de 5-clique, mas não 5-clãs porque o diâmetro é igual a sete. Os grafos "brock" têm diâmetro igual a 2 e uma cobertura por um 2-clique é suficiente. A maior parte dos casos dos ficheiros DIMACS apresenta este perfil. Por outro lado, os grafos "c-fat", têm diâmetros maiores que 7, gerando longas sequências de coberturas de k-cliques. Na medida proposta, a sequência de cobertura por k-cliques, identifica famílias de grafos, e parece ser bastante promissora.

4 Conclusões

Dada a grande quantidade de dados, fornecidos pela Web 2.0, existe uma necessidade permanente de criar novas medidas para melhor compreender a estrutura das redes, o modo como as suas componentes estão organizados e o modo como evoluem ao longo do tempo.

As medidas na análise de Redes Complexas são essencialmente baseadas em procedimentos de baixa complexidade computacional, como o diâmetro do grafo, a distribuição do grau dos nós e a verificação da conectividade, subestimando o conhecimento da estrutura das componentes do grafo.

Neste trabalho o conceito de clique-relaxado é estendido a todo o grafo, quando este é coberto por k-cliques, obtendo assim uma visão geral das sub-componentes do grafo. A sequência de coberturas por k-cliques é apresentada, combinando a análise de subgrupos coesos com o conceito dos "pequenos mundos". A análise das sequências identifica diferentes tipos de grafos, mostrando que famílias de grafos têm estruturas diferentes.

Existe ainda um conjunto de características, não mencionadas no presente documento, mas que podem ser obtidas, como a sobre-coberturas dos nós, a cardinalidade das k-cliques e a composição das k-cliques.