Detecção de Comunidades em Redes Complexas com informações associadas às arestas

Steven Koiti Tsukamoto
Orientação: Prof. Marcelo Finger
Ana Paula Appel, PhD

Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo

Resumo: O surgimento e crescimento de dados modelados como grafos, como as redes complexas, tornou possível a criação de novas áreas de pesquisa como a mineração de grafos. Nesta área as tarefas que se destacam são a extração de propriedades estatísticas, a detecção de grupos (comunidades) e a predição de ligações (arestas).

A detecção de comunidade pode ser utilizada para pré-processamento do grafo afim de quebrá-lo em grupos que podem ter suas propriedades mineradas mais facilmente por serem menores. Contudo, a grande maioria dos métodos de detecção de comunidades funciona para grafos não direcionados que não possuem informações associadas a nós e arestas e que possuem apenas uma aresta entre os nós.

O principal estudo de caso desse projeto é o projeto Leitura da Web, que visa à construção de uma ontologia, que pode ser modelada como uma rede complexa, para representar o conhecimento extraído da Web por um algoritmo de aprendizado sem fim. O grafo produzido pela extração do conhecimento da Web possui informações associadas às arestas e pode em alguns casos ter mais de uma aresta associada entre os pares de nós.

Assim, o objetivo deste projeto é estudar os algoritmos de detecção de comunidade que se apliquem a esse grafo e disponibilizá-los no framework Centauro que está sendo construído.

Palavras chave: mineração redes complexas, detecção de comunidades.

1 Introdução

O grande volume de dados nos últimos anos impulsionou o crescimento da área de descoberta de conhecimento em grandes bases de dados, com especial atenção ao desenvolvimento de algoritmos eficientes para a mineração de dados.

Contudo, não só o volume de dados tem aumentado, mas também a sua variedade. Um exemplo são as redes complexas, que têm sido usadas cada vez mais frequentemente em diversos domínios de aplicação. Dentre esses domínios, podemos destacar o biológico (cadeia de proteínas e DNA), relações sociais (redes sociais/relacionamento tais como Orkut, Facebook, Myspace, etc.), redes ponto a ponto (edonkey, gnutella, etc), relacionamento acadêmico (DBLP, ARXIV, wikipedia).

O estudo das redes complexas tem proporcionado a descoberta de padrões interessantes, como a distribuição do grau dos nós [Albert et al., 1999, Adamic et al., 2000], diminuição do diâmetro das redes com a sua evolução [Leskovec et al., 2005], o fenômeno do "Small World", na qual todo mundo esta

ligado a todo mundo em poucos passos [Milgram, 1967], entre outros. Esses padrões ajudam a entender não só o relacionamento social entre os seres humanos com as redes sociais [Leskovec et al., 2008], mas também a disseminação de informações e doenças [Chakrabarti et al., 2008], detecção de intrusão em redes de computadores e sensores e formação de comunidades e grupos em redes [Fortunato, 2010]. Uma boa visão da área é apresentada em [Newman, 2003].

A base de conhecimento que está sendo criada e continuamente expandida pelo projeto Leitura da Web (<http://rtw.ml.cmu.edu/rtw/>) é representada por uma ontologia, que é um modelo de dados que representa um conjunto de conceitos dentro de um domínio e os relacionamentos entre estes.

Uma ontologia pode ser representada como uma rede complexa e assim, modelada como um grafo.

Um dos objetivos do Framework Centauro [Appel and Hruschka Jr., 2010] é permitir que os usuários experimentem diversas combinações de algoritmos, sejam eles para executar a mesma tarefa ou não. Assim o primeiro passo deste projeto é fazer uma ampla análise dos algoritmos de detecção de comunidades que mais se adequam a estrutura do grafo produzido pela base de conhecimento do projeto Leitura da Web.

2 Objetivos

O objetivo da área de mineração de redes complexas pode ser descrito, em poucas palavras, como prover soluções eficientes para a descoberta de padrões em dados modelados como grafos.

A detecção de grupos, comunidades, em redes complexas tem por objetivo definir grupos de nós que possuem mais conexões entre si do que com o restante da rede [Fortunato, 2010]. Assim, a detecção de comunidade pode ser uma ferramenta importante para o particionamento eficiente do grafo para a posterior mineração ou armazenamento do mesmo.

A Web pode ser considerada uma fonte inesgotável de dados que podem ser modelados como grafo e assim tratada como uma rede complexa. Neste contexto está inserido o projeto Leitura da Web, que tem por objetivo o desenvolvimento de métodos e algoritmos que permitam a construção de um sistema computacional capaz de realizar a extração de conhecimento a partir da Web em português e em inglês, por meio da criação de uma base de conhecimento consistente, que seja atualizada constantemente à medida que novos conhecimentos vão sendo extraídos.

Como pode ser visto em [Betteridge et al., 2009], [Tom M. Mitchell, 2009], o principal objetivo do projeto Leitura da Web é a construção de um sistema computacional de aprendizado sem fim, e a ideia chave é a integração de diversos métodos e técnicas de extração de conhecimento que viabilize a evolução automática do sistema e sua ontologia.

A ontologia inicial a ser fornecida ao sistema deve especificar um conjunto de categorias (por exemplo, cidade, país, estado, ator, político, empresa, universidade, etc.) e relações (por exemplo, prefeito(político, cidade), capitalDoPaís(cidade, país), capitalDoEstado(cidade, estado), etc.) a serem utilizadas como guia na extração do conhecimento e que serão populadas com instâncias encontradas na Web.

A ontologia será constantemente atualizada pelo sistema e formará uma base conhecimento dinâmica que, por meio de um motor de inferência, será utilizada para prover respostas a perguntas feitas pelos usuários.

5 Referências

- [Adamic et al., 2000] Adamic, L. A., Huberman, B. A., Barabási, A., Albert, R., Jeong, H., and Bianconi, G. (2000). Power-law distribution of the world wide web. *Science*, 287(5461):2115a+.
- [Albert et al., 1999] Albert, R., Jeong, H., and Barabási, A.-L. (1999). The diameter of the world wide web.
- [Appel and Hruschka Jr., 2010] Appel, A. P. and Hruschka Jr., E. R. (2010). Centaurs a component based framework to mine large graphs. In *XXV BRAZILIAN SYMPOSIUM ON DATABASES*, pages 1–8, Belo Horizonte, MG, Brazil.
- [Betteridge et al., 2009] Betteridge, J., Carlson, A., Hong, S. A., Hruschka Jr., E. R., Law, E. L. M., Mitchell, T. M., and H.Wang, S. (2009). Toward never ending language learning. In *AAAI Spring Symposium*.
- [Chakrabarti et al., 2008] Chakrabarti, D., Wang, Y., Wang, C., Leskovec, J., and Faloutsos, C. (2008). Epidemic thresholds in real networks. *ACM Trans. Inf. Syst. Secur.*, 10(4):1–26.
- [Clauset et al., 2008] Clauset, A., Moore, C., and Newman, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101.
- [Fortunato, 2010] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.
- [Leskovec et al., 2008] Leskovec, J., Backstrom, L., Kumar, R., and Tomkins, A. (2008). Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, New York, NY, USA. ACM.
- [Leskovec et al., 2005] Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *eleventh ACM SIGKDD*, pages 177–187, New York, NY, USA. ACM Press.
- [Liben-Nowell and Kleinberg, 2003] Liben-Nowell, D. and Kleinberg, J. (2003). The link prediction problem for social networks. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, New York, NY, USA. ACM.
- [Milgram, 1967] Milgram, S. (1967). The small world problem. *Psychology Today*, 2:60–67.
- [Newman, 2003] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45:167–256.

[Taskar et al., 2004] Taskar, B., Wong, M., Abbeel, P., and Koller, D. (2004). Link prediction in relational data.

[Tom M. Mitchell, 2009] Tom M. Mitchell, Justin Betteridge, A. C. E. R. H. R. W. (2009). Populating the semantic web by macro-reading internet text. In *Proceedings of the 8th ISWC*.