DocGenealogy
Uma árvore genealógica de doutorados

David Fernandes
Universidade Aberta
Rua da Escola Politécnica 141-147, 1269-001 Lisboa
Portugal
david.paiva.fernandes@gmail.com

Elizabeth Carvalho
CIAC – Centro de Investigação em Artes e Comunicação
Rua da Escola Politécnica 141-147, 1269-001 Lisboa
Portugal
ecarvalho@uab.pt

1. Introdução

O estudo de actividades complexas, tais como a produção científica e o desenvolvimento de software requerem muitas vezes a modelagem de conexões entre entidades heterogéneas, incluindo artefactos (Heer & Perer, 2014), pessoas e instituições. **DocGenealogy** é uma aplicação web que permite facilmente descobrir a relação existente entre orientadores de doutoramento e seus orientados ao longo do tempo. Ela oferece um gráfico interactivo para ajudar nesta tarefa, além de outros gráficos para suporte a uma melhor análise e entendimento da informação pelos utilizadores finais.

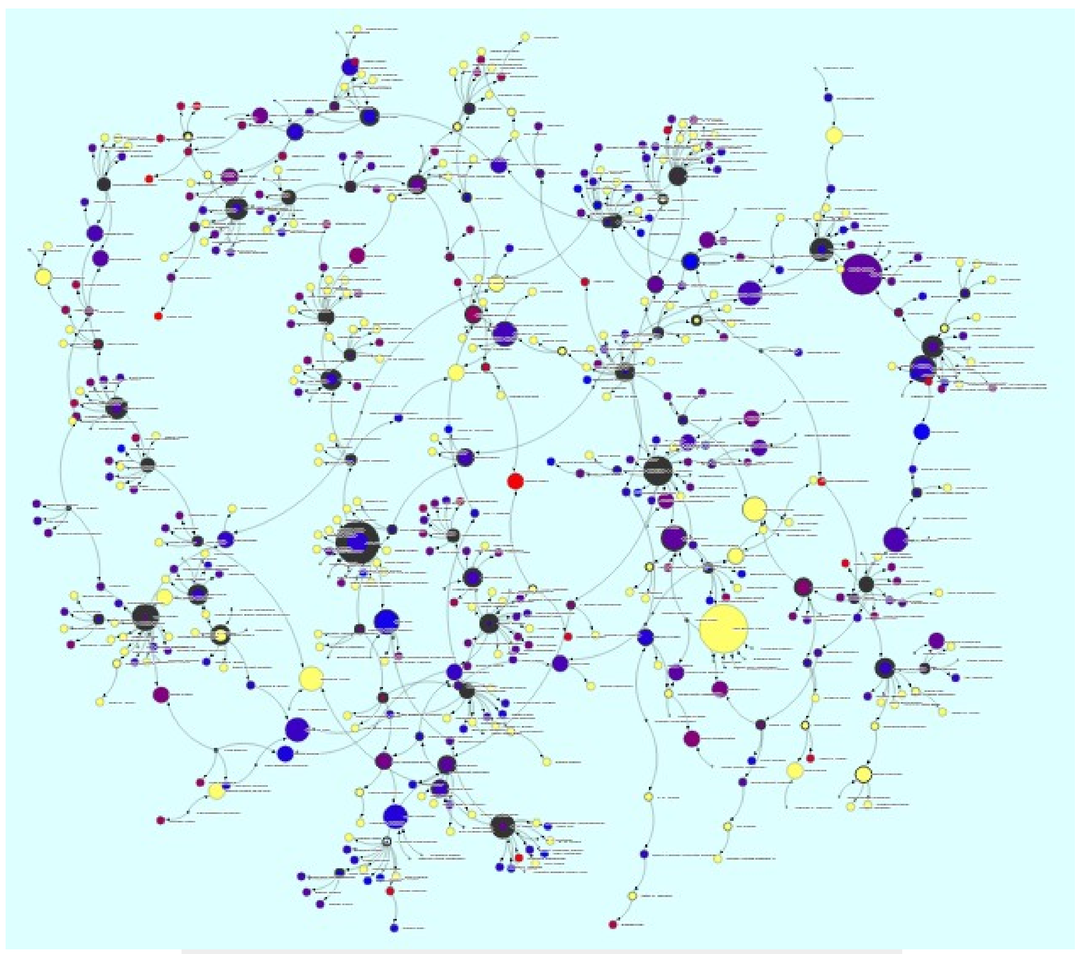
2. Base científica

Grafos

De uma forma geral, os grafos são um dos mais importantes modelos de dados em ciências da computação, porque muitos problemas e domínios podem ser modelados como estruturas desse tipo.

Os cientistas criaram uma diversidade de técnicas de visualização para grafos: **diagramas vértice-arco** (Jianu, Rusu, Hu, & Taggart, 2014), **vistas de matrizes de adjacência** (Kang, Lee, Koutra & Faloutsos, 2014), **híbridos das duas foram também propostos** (Rufiange, McGuffin & Fuhrman, 2012; Henry, Fekete & McGuffin, 2007).

Shneiderman e Aris (2006) propuseram um *layout* de visualização em rede com base em substrato semântico definido pelo utilizador com o diagrama de ligações dos vértices como uma visualização subjacente.



http://wst.life/docgenealogy

3. Material e método

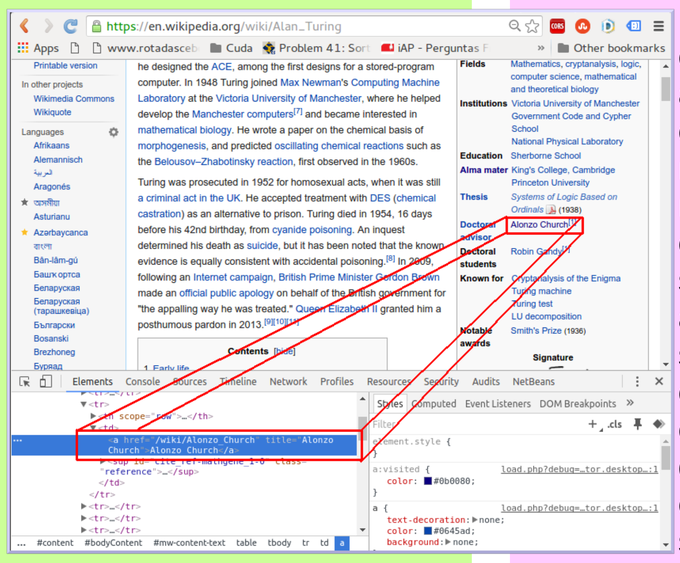
Origem dos dados

Infobox "cientista"

Um *infobox* de um artigo na Wikipédia (Tran & Cao, 2013) contém geralmente os factos-chave e é organizado como pares de atributo-valor. Os *infoboxes* permitem não só aos leitores reunirem rapidamente as informações mais importantes sobre alguns aspectos dos artigos em que eles aparecem, mas também fornecem uma fonte para muitas bases de conhecimento derivadas da *Wikipedia*. No entanto, nem todos os valores dos atributos de uma *infobox* são actualizados com frequência e com precisão.

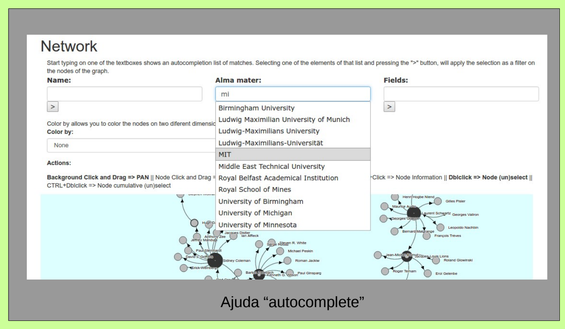
Extracção de dados

O resultado e as informações a que temos acesso, como em qualquer outro tipo de página web dinâmica, é em HTML puro. A ferramenta de extracção de dados que implementamos, lê o HTML de uma página, extrai as informações relevantes necessárias e, fazendo uso dos URL de orientadores/mentores de doutoramento e de orientados, recursivamente obtém mais informações, construindo assim o grafo.



Interface e visualização

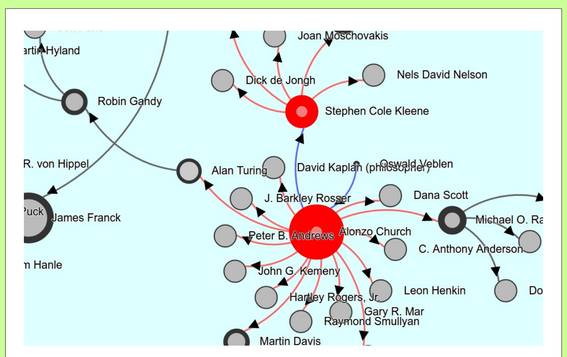
Os algoritmos de força-direta (Kobourov, 2012) estão entre os métodos mais flexíveis para o cálculo de *layouts* de grafos simples, sem orientação. Também conhecido como molas encastradas (*spring embedders*), tais algoritmos calculam o *layout* de um grafo utilizando apenas informações contidas dentro da estrutura do grafo propriamente dito, em vez de depender do conhecimento do seu domínio específico. Os grafos desenhados com estes algoritmos tendem a ser esteticamente agradáveis, exibem simetrias e tendem a produzir *layouts* sem cruzamentos entre arcos, livres para os grafos planares.



Ajuda “autocomplete”

Tecnologia

Para a extracção dos dados e o processo de rastreamento *web* foi implementada uma aplicação *java* específica que executa as solicitações directas via *http*. O conteúdo HTML de cada página é explorado utilizando consultas XPath XML (W3C, 2015), que extraem os dados relevantes e organiza-os segundo estruturas de dados JSON. Nesse formato, os dados ficam prontos para o consumo directo pela *framework* de visualização e interação gráfica, D3.js (Bostock, 2016a), uma biblioteca em JavaScript que utiliza o HTML, SVG e CSS, em especial, as suas funcionalidades de desenho de grafos (Bostock, 2016b). Finalmente, foi criado (http://wst.life/docgenealogy) um *website* para mostrar a visualização e a interacção disponíveis sobre os dados, utilizando padrões em HTML/CSS/JS.



Selecção de vértices e visualização de relações

4. Trabalho futuro e considerações

Este artigo introduziu o aplicativo web **DocGenealogy,** um trabalho em curso. Já realizamos alguns testes informais e preliminares para avaliar seu desempenho e sua qualidade visual total, tanto a nível de estética, além da usabilidade de sua interface e eficácia no mapeamento de dados. Apesar de termos tido um resultado global muito positivo, como o nosso objectivo principal é conhecer a sua eficácia para apoiar a análise visual e raciocínio sobre a relação entre orientadores de doutoramento e orientados, estamos desenvolvendo actualmente um estudo de caso, que permita a sua avaliação e teste de forma mais ampla e detalhada em termos de sua eficácia para apoiar a apreensão da informação.

Dados originais e ligação à página da Wikipedia

Referências:

Bostock, M. (2016a). Data-driven documents. https://d3js.org/. Accessed: 2016-1- 15.

Bostock, M. (2016b). Force layout. https://github.com/mbostock/d3/wiki/Force-Layout. Accessed: 2016-1- 15.

Beck, F., Burch, M., Diehl, S., & Weiskopf, D. (2014). The state of the art in visualizing dynamic graphs. EuroVis STAR.

Dunne, C., & Shneiderman, B. (2013, April). Motif simplification: improving network visualization readability with fan, connector, and clique glyphs. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 3247-3256). ACM.

Heer, J., & Perer, A. (2014). Orion: A system for modeling, transformation and visualization of multidimensional heterogeneous networks. Information Visualization, 13(2), 111-133.

Henry, N., Fekete, J. D., & McGuffin, M. J. (2007). NodeTrix: a hybrid visualization of social networks. Visualization and Computer Graphics, IEEE Transactions on, 13(6), 1302-1309.

Jianu, R., Rusu, A., Hu, Y., & Taggart, D. (2014). How to display group information on node-link diagrams: an evaluation. Visualization and Computer Graphics, IEEE Transactions on, 20(11), 1530-1541.

Kang, U., Lee, J. Y., Koutra, D., & Faloutsos, C. (2014). Net-ray: Visualizing and mining billion-scale graphs. In Advances in Knowledge Discovery and Data Mining (pp. 348-361). Springer International Publishing.

Kobourov, S. G. (2012). Spring Embedders and force directed Graph Drawing Algorithms, university of Arizona.

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., & Carpendale, S. (2012). Empirical studies in information visualization: Seven scenarios. Visualization and Computer Graphics, IEEE Transactions on, 18(9), 1520-1536.

Liu, H., & Motoda, H. (2012). Feature selection for knowledge discovery and data mining (Vol. 454). Springer Science & Business Media.

Liu, S., Cui, W., Wu, Y., & Liu, M. (2014). A survey on information visualization: recent advances and challenges. The Visual Computer, 30(12), 1373-1393.

Munzner, T. (2014). Visualization Analysis and Design. CRC Press.

Rufiange, S., McGuffin, M. J., & Fuhrman, C. P. (2012, February). TreeMatrix: A hybrid visualization of compound graphs. In Computer Graphics Forum (Vol. 31, No. 1, pp. 89-101). Blackwell Publishing Ltd.

Shneiderman, B., & Aris, A. (2006). Network visualization by semantic substrates. Visualization and Computer Graphics, IEEE Transactions on, 12(5), 733-740.

Tran, T., & Cao, T. H. (2013). Automatic Detection of Outdated Information in Wikipedia Infoboxes. Research in Computing Science, 70, 183-194.

W3C (2015. XML Path Language (XPath). Accessed: 2016-1-15.

Zhou, H., Xu, P., Yuan, X., & Qu, H. (2013). Edge bundling in information visualization. Tsinghua Science and Technology, 18(2), 145-156.

Zinsmaier, M., Brandes, U., Deussen, O., & Strobelt, H. (2012). Interactive level-of-detail rendering of large graphs. Visualization and Computer Graphics, IEEE Transactions on, 18(12), 2486-2495.