Analítica Visual (Visual Analytics) de Topologia e Texto em Redes Sociais

Proponente: Renato Fabbri
Supervisora: Profa. Dra. Maria Cristina Ferreira de Oliveira
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo (ICMC-USP)

27 de abril de 2017

Resumo:
Palavras-chave: Visual analytics, Redes sociais, Redes complexas, Mineração de texto.

Analítica Visual (Visual Analytics) de Topologia e Texto em Redes Sociais

Proponent: Renato Fabbri
Supervisor: Prof. Dr. Maria Cristina Ferreira de Oliveira
Institute of Mathematics and Computational Sciences,
University of São Paulo (ICMC-USP)

27 de abril de 2017

Abstract
Keywords: Visual analytics, Social networks, Complex networks, Text mining.

1 Enunciado do problema

Este trabalho se propõe a desenvolver métodos e ferramentas para a Análise de Redes Sociais (ARS) via analítica visual (*visual analytics*) e considerando as propriedades das estruturas sociais. Exemplos destas propriedades podem ser divididos em três grupos:

- Topoógicos: as redes sociais são reportadas como livres de escala, de pequeno mundo e com comunidades [**?**].
- Textuais: a linguagem nas redes sociais é muitas vezes informal, com abreviações e erros com relação às normas oficiais [**?**].
- Relacionamento entre topologia e texto: há traços do texto produzido pelos participantes que se relacionam à topologia. Hubs utilizam frases mais curtas e mais verbos e advérbios, periféricos utilizam mais substantivos [**?**].

Em resumo, nossa proposta de desenvolver uma ferramenta de analítica visual para redes sociais envolve o processamento dos dados para obtenção de estruturas relevantes, da concepção de visualizações e formas de interação com elas, da integração dos dados a serem analisados e da escrita do software para processamento dos dados e para a interface de utilização para o usuário.

1.1 Sua importância

Há uma proliferação de redes sociais virtuais que produzem muitos dados e estas estruturas estão ainda sendo caracterizadas. Uma ferramenta de analítica visual em cógido aberto, portável para sistemas operacionais diferentes, e que seja capaz de lidar com grandes quantidades de dados é certamente algo desejável. As análises a serem possibilitadas pela ferramenta, e necessárias para a concepção da mesma, são úteis para aprofundar nossa compreensão das redes sociais virtuais.

1.2 A contribuição se bem sucedido

Do ponto de vista dos métodos, as análises pertinentes, assim como escolhas de interfaces apropriadas, já resultaram em publicação científica pelo proponente [**?**] e pela supervisora [**?**, **?**] e deverá implicar em maiores contribuições para a compreensão da analítica visual e das estruturas sociais virtuais. O software a ser desenvolvido certamente é uma contribuição central pois possibilitará avanços na analítica visual e na caracterização dos sistemas.

1.3 Trabalhos relevantes

A tese de doutoramento do proponente *Topologial stability and textual differentiation in human interaction networks: statistical analysis, visualization and linked data* apresenta contribuições para a caracterização dos sistemas sociais virtuais, algumas destas em processo de publicação [**?**, **?**, **?**] e uma já aceita pela

revista Physica A [?]. Há trabalhos do grupo de pesquisa em visualização de dados e de redes [?, ?, ?]. Os trabalhos [?, ?, ?] são base para visualização de dados, enquanto [?, ?, ?] são especificamente para visualização de redes/grafos e [?, ?, ?] são sobre visualização para mineração de texto. Os trabalhos [?, ?, ?] são bastante relevantes sobre visual analytics e [?, ?] são referência sobre interação humano-computador. Há diversos software que deverão ser utilizados no decorrer do trabalho para apreender as possibilidades de analítica visual neste contexto (e.g. Gephi, Weka, ccNetViz, D3.js, Three.js). Há também literatura e artefatos mais ligados à arte, como a teoria da Gestalt, kiki-bouba, design e estética.

1.3.1 Projeto temático

A supervisora desta proposta, a Profa. Dra. Maria Cristina Ferreira de Oliveira, está pesquisando e possui um aluno de mestrado no tema. Ela enviou recentemente uma proposta de projeto temático para a FAPESP. Segue o resumo.

"A pesquisa em Visual Analytics é central no tratamento dos desafios associados à análise de dados e computação intensiva em dados, pelo potencial de combinar técnicas de Aprendizado de Máquina e de Visualização para apoiar a interpretação de dados complexos. O acoplamento de técnicas oriundas de ambas as áreas pode promover avanços significativos na capacidade humana de análise de dados, pois permite a indivíduo e computador assumirem papeis complementares ao tratar os muitos problemas introduzidos pelo volume e complexidade dos conjuntos de dados gerados em diversos domínios de aplicação. Este projeto de pesquisa aborda dois focos distintos em visual analytics, um de natureza aplicada e outro de natureza conceitual. No aspecto aplicado serão considerados (i) o problema de visualização de redes de grande escala, com ênfase em redes sociais; e (ii) o problema da análise exploratória de espaços de atributos que caracterizam fenômenos multivariados e variantes no tempo - por exemplo, resultantes de sensores utilizados para monito- ramento ambiental em diversos domínios. Em ambos os casos, a busca por soluções escaláveis para grandes volumes de dados representa um desafio. No aspecto con- ceitual, dando continuidade a uma colaboração em andamento, iremos conceber e realizar alguns estudos experimentais que contribuam para esclarecer os processos cognitivos subjacentes à interpretação de um tipo particular de visualização mul- tidimensional, os chamados mapas de similaridade. A análise dos resultados pode sugerir modelos conceituais sobre a interpretação desse tipo de mapeamento visual. Esperamos com esse estudo contribuir para ampliar o embasamento conceitual sobre essas técnicas, essencial para futuros avanços na área."

2 Resultados esperados

São esperados resultados em três frentes:

- Melhor caracterização das redes sociais virtuais através de mineração da topologia e do texto.

- Avanços em analítica visual de redes sociais principalmente através da consideração das características topológicas e textuais para a interface.
- Um software aberto para analítica visual de redes sociais que seja capaz de lidar com grandes volumes de dados e seja executado em navegadores web.

2.1 Disseminação dos resultados

O software aberto pretendido deve facilitar a interação com usuários, desenvolvedores e pesquisadores. Pretendemos apresentar resultados em conferências e em artigos de revista. A circulação de material didático em video deve aumentar a recepção das contribuições. As análises e interfaces devem ser devolvidas às comunidades virtuais que geraram os dados.

3 Desafios científicos e tecnológicos

A consideração da tarefa revela alguns desafios. Primeiro, a análise de texto e de topologia implicam em um conjunto de medidas de alta dimensionalidade. Segundo, o montante de dados a serem analisados é bastante grande e as medições são muitas vezes computacionalmente caras. Por exemplo, a medida usual de *betweenness centrality* requer que sejam encontradas todas as geodésicas entre todos os pares de vértices [**?**]; as medidas de texto são resultantes de muitos dados e podem envolver procedimentos não triviais como a etiquetação morfossintática ou similaridades via rede semântica (Wordnet) [**?**].

Este projeto propõe a utilização destas análises, e aprofundamento delas, para a analítica visual. Os desafios imediatos para esta tarefa são:

- O uso apropriado de técnicas de visualização. Os layouts para redes orientados a força são os mais usuais em redes sociais por explicitarem comunidades e resultarem em figuras esteticamente atraentes, mas a observação de outras características ficam prejudicadas, um fenômeno que tem recebido o nome de *hairball effect*. Para a visualização de dados multidimensionais, as técnicas podem não ser apropriadas para dados com dimensionalidade muito alta ou não resultarem em coordenadas com fácil interpretação (e.g. no caso do LSP). Além disso, o uso e.g. de cores, formas e glifos possuem um alto impacto na capacidade de análise pelo usuário.
- Viabilizar a interatividade com o montante de dados e medições exposto acima. Quais controles devem ser disponibilizados e com quais controladores? Qual o melhor balanço entre interatividade e quantidade de dados a serem analisados?
- Possibilitar a análise de estruturas arbitrárias.
- A escrita de software.

3.1 Meios e métodos para superá-los

Os desafios da análise de topologia e texto em redes sociais foram abordados pelo proponente em [**?**], resultando em caracterização das estruturas sociais quanto à estabilidade topológica e diferenciação textual. Para os outros desafios ligados explicitamente á analítica visual, itemizados acima, propomos:

- A utilização de layouts geometricamente inspirados, como os *Hive Plots* [**?**] ou o Versinus (este desenvolvido pelo prroponente [**?**]). A possibilidade de utilização de layouts e plots diferentes e de layouts simulâneos em regiões distintas da interface.

- O estudo da literatura da área pode nos fornecer pistas valiosas, mas certamente só saberemos quais as reais possibilidades de interatividade através da implementação e de testes subsequentes. Prevemos também que, evidentemente, as possibilidades de interatividade de processamento massivo de dados dependerá do hardware do usuário, uma vez que deverá ser executado no cliente ao menos a visualização.

- Para que as estruturas a serem analisadas possam ser integradas e observadas por partes, podemos aproveitas a representação de estruturas sociais como dados ligados já desenvolvida pelo proponente [**?**, **?**]. O banco de dados resultante preferencialmente deve ser abrigado em alguma instancia publicamente acessível e pode contar com expansões na medida em que seja conveniente.

- Entendemos que o ideal é que o software possa ser executado em navegadores web usuais (e.g. Google Chrome e Firefox). Para isso, contamos com boas bibliotecas de visualização. A principal é a D3.js, que pode ser integrada com WebGL para possiblitar visualizações mais complexas. Boas alternativas são o Three.js e o ccNetViz, que já são integrados com WebGL. Já no processamento dos dados, podemos aproveitar o bom repertório de bibliotecas em Python, que utiliza as rotinas em BLAS e LAPACK para processamento matricial e vetorial. Devemos estudar as possiblidades de implementar o processamento dos dados diretamente em JavaScript para que a arquitetura do software fique simplificada, mas isso não é central, dado que a integração de JavaScript com Python é imediato caso sejam utilizados frameworks em Python para desenvolvimento web como Flask e Django. A maior diferença neste caso é que a implementação das rotinas de processamento dos dados em Python serão executadas pelo servidor ao passo que, se implementadas em JavaScript, poderão ser executadas no cliente (caso seja feita uma implementação em Node.js, o processamento em JavaScript poderá ser executado no servidor). Concebemos toda a implementação em código aberto/software livre para facilitar a colaboração no desenvolvimento e permitir que as rotinas sejam escrutinizadas.

4 Cronograma

24 meses.

5 Disseminação e avaliação

Como exposto na Seção 2.1, haverá disseminação das contribuições em conferências, revistas, video e através do software em código aberto. A avaliação deverá ser feita em três frentes: pela nossa própria capacidade de mineração dos dados, idealmente para obtenção de resultados suficientemente relevantes para publicação; através de testes com usuários em potencial; através da submissão das contribuições de analítica visual para revistas, de forma a obter avaliações de revisores especialistas.

6 Outros apoios

FAPESP, grupo, GSoC?.