Extração de Conhecimento e Análise Visual de Redes Sociais

Carla M. D. S. Freitas, Luciana P. Nedel, Renata Galante, Luís C. Lamb, André S. Spritzer, Sérgio Fujii, José Palazzo M. de Oliveira, Ricardo M. Araújo, Mirella M. Moro

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil
{carla,nedel,galante,lamb,spritzer,syfujii,palazzo,rmaraujo,mirella}@inf.ufrgs.br

Abstract. *A social network is a graph where people or organizations (depending on the application) are represented as nodes connected by edges that can refer to either tight social bonds or some common, shared aspect. The graph structure analysis and the statistical analysis of specific node/edge attributes can reveal important individuals, relationships, and clusters. New information continues to be collected and stored, and size and complexity of the semantic graphs overwhelm the human cognitive abilities. Hence, it is necessary to improve the computational mechanisms to analyze such volume of data. In this paper, we focus on analyzing the information from social networks, extracting relevant knowledge, and visualizing the facts resultant from the analysis.*

Resumo. *Uma rede social é um grafo onde pessoas ou organizações (dependendo da aplicação) são representadas por nodos conectados por arestas as quais podem corresponder tanto a fortes relacionamentos sociais como ao compartilhamento de alguma característica. A análise da estrutura desse grafo, assim como a análise estatística dos atributos dos nodos e/ou das arestas pode revelar indivíduos/organizações importantes, relacionamentos especiais e grupos. Enquanto novas informações continuam a ser coletadas e armazenadas, e o tamanho e a complexidade dos grafos semânticos sobrepujam a capacidade cognitiva humana, é necessário melhorar a habilidade de analisar tais volumes de dados. Este artigo focaliza a análise da informação presente nas redes sociais, a extração de conhecimento a partir de grafos e a visualização de fatos decorrentes dessa análise.*

1. Introdução

Redes sociais tornaram-se especialmente relevantes devido à grande variedade de sites Web que utilizam o conceito, como Orkut, MySpace, FaceBook e Flickr. Seus usuários formam bases de dados que provêem um importante meio de compartilhar, organizar e encontrar conteúdo, contatos e estabelecer interesses comuns. Devido ao uso intenso, estes sites reúnem material suficiente para subsidiar estudos de características de redes sociais em larga escala. Kumar et al. (2006b) mostraram que o poder de interação entre as pessoas pode ser a razão da falha ou do sucesso de uma organização.

A análise e a extração de conhecimento de redes sociais vêm sendo amplamente utilizadas em várias áreas, incluindo ciências sociais e comportamentais, economia e

marketing [Wasserman 1994], onde a compreensão do comportamento da sociedade é estratégica. A análise do conhecimento explícito com o objetivo de adquirir melhores condições de competitividade, denomina-se inteligência competitiva. Esta área de pesquisa conta com um corpo de conhecimentos próprios, além de inúmeras contribuições de outras áreas tais como marketing, *benchmarking* e análise estratégica.

A maior parte das informações concorrenciais que os profissionais de inteligência competitiva utilizam são públicas, o que facilita o acesso e a busca pelas organizações. Day e Wensley (1988) propõem três condições fundamentais a serem observadas como determinantes de vantagem competitiva: a primeira considera como vantagem a integração das fontes componentes, da posição e do resultado do desempenho da organização. A segunda mostra que o conceito de vantagem reside em uma habilidade superior, na disponibilidade de recursos que são revelados através da competitividade do produto da organização no mercado. Um ponto de vantagem pode ser proveitoso somente quando oferece benefícios que são percebidos e valorizados pelo cliente, e que são difíceis par o concorrente oferecer. Finalmente, a terceira identifica os produtos e os mercados para os quais a organização está realmente capacitada para atuar. Esta vantagem surge do valor que a organização consegue criar e o que é percebido pela sociedade. No caso de Universidades, este produto é a competência dos grupos de pesquisa.

Métodos tradicionais de aprendizagem de máquina e mineração de dados geralmente recebem como entrada um conjunto randômico de dados homogêneos para os quais é computada uma saída. Em geral, os dados de redes sociais são heterogêneos, multi-relacionais e semi-estruturados. Estudos visando explorar a descoberta de conhecimento em redes sociais envolvem áreas diversas como análise de *links*, mineração de dados, mineração de grafos, aprendizagem de máquina e técnicas de visualização. Especificamente, o princípio fundamental na visualização de redes sociais é facilitar a compreensão dos dados. As técnicas existentes direcionam-se a selecionar sub-grupos de informações com o objetivo de simplificar a visualização. Todavia, os métodos empregados geralmente envolvem a seleção manual de dados, como proposto por Henry e Fekete (2007), ou de simples análise em que é verificado o casamento de atributos entre as entidades [Wasserman 1994].

O objetivo deste artigo é apresentar uma visão geral dos problemas envolvidos na área de descoberta de conhecimento e visualização de informações em redes sociais. Deste modo, o artigo contempla a especificação de um processo para análise de conhecimento e visualização de dados de redes sociais. Para análise de conhecimento é especificada uma estrutura para representação dos dados, um conjunto de técnicas de mineração de dados para aquisição de conhecimento e um mecanismo de aprendizagem de máquina para atuarem racionalmente no ambiente e incrementar o desempenho em tarefas futuras. São discutidos também aspectos relativos à visualização que permitem de maneira gráfica e interativa explorar as redes sociais, bem como analisar o conhecimento adquirido durante o processo de mineração e aquisição de conhecimento. Por fim, é apresentado um estudo de caso que mostra possibilidades de análise de uma rede social de colaboração científica entre pesquisadores com dados provenientes das bibliotecas digitais BDBComp [Laender 2004] e DBLP (*Digital Bibliography & Library Project*) [DBLP 2007]. A motivação para este estudo de caso é que o

conhecimento das redes de colaboração científica é elemento essencial para a avaliação da qualidade, através do nível de inserção científica, dos grupos de pesquisa.

Como pôde ser observado pelos problemas levantados acima em relação à extração de informações de redes sociais, os mesmos estão fortemente relacionados à gestão da informação em grandes volumes de dados distribuídos, que constitui o primeiro dos grandes desafios da pesquisa em computação no Brasil, identificados por pesquisadores da SBC em 2006.

O restante do texto está organizado da seguinte forma. A Seção 2 introduz o conceito de rede sociais. A Seção 3 especifica detalhadamente o processo de análise de conhecimento e da visualização propostos, enquanto a Seção 4 discute tal processo através de um estudo de caso de análise visual de redes sociais de colaboração científica entre pesquisadores. A Seção 5 apresenta cenários, aplicações e desafios na área de análise de conhecimento e visualização de redes sociais. Finalmente, a Seção 6 discute os resultados preliminares obtidos no âmbito dos Grandes Desafios da SBC e apresenta perspectivas de trabalhos futuros.

2. Redes Sociais

Redes sociais são representadas por grafos onde os nodos são os atores (geralmente pessoas) e as arestas são os relacionamentos entre esses atores. Podem apresentar desde conexões esparsas, como em árvores genealógicas, até conexões muito densas, como em redes de contato na internet. As redes sociais são normalmente classificadas em três categorias: redes aleatórias, redes de mundos pequenos e redes livres de escala. Erdös e Rényi (1960) propuseram uma teoria sobre redes aleatórias. Analisando a construção de redes sociais, demonstraram que bastava uma conexão entre cada um dos convidados de uma festa para que todos estivessem conectados ao final dela. Adicionalmente, verificaram que, quanto mais conexões são adicionadas, maior é a probabilidade de serem gerados clusters. Portanto, uma festa poderia ser definida através de um conjunto de clusters que, de tempos em tempos, estabelecem relações aleatórias com outros grupos. Assim, concluíram que todos os nodos, em uma determinada rede, teriam aproximadamente a mesma quantidade de conexões, constituindo-se em redes igualitárias [Barabasi 2003].

Na década de 1960, Milgram realizou um experimento para observar o grau de separação entre as pessoas [Degenne 1999]. Ele enviou aleatoriamente uma quantidade de cartas a vários indivíduos, solicitando que tentassem redirecioná-las a um alvo específico. Se não conhecessem o alvo, as pessoas eram solicitadas a enviar as cartas para alguém que acreditassem estar mais próxima a ele. Milgram descobriu que, das cartas que chegaram a seu destinatário final, a maioria havia passado apenas por um pequeno número de pessoas. Isso sugeriria que todas estariam a poucos graus de separação umas das outras, ou seja, em um “mundo pequeno”. Este estudo é chamado de fenômeno do mundo pequeno (*small-world phenomenon*), também conhecido como “princípio dos seis graus de separação”, onde se pretende provar que cada ator está conectado a qualquer outro na rede com um número máximo de seis atores intermediários.

Apesar de estabelecer certos padrões, Milgram, e posteriormente Watts (2003), tratavam as redes sociais como redes aleatórias, ou seja, redes em que as conexões entre os nós eram estabelecidas de modo randômico, exatamente como Erdös e Rényi. Por

sua vez, Barabási (2003) demonstrou que as redes não são formadas de modo aleatório, mas que existe uma ordem na dinâmica de estruturação das redes. Este padrão de estruturação foi identificado por Barabási e aponta para o fato de que "ricos ficam mais ricos" (*rich get richer*), ou seja, quanto mais conexões um nó possui, maiores as chances de ele ter mais novas conexões. Ele chamou essa característica de conexão preferencial: um novo nó tende a se conectar com um nó pré-existente, mas mais conectado. Isso implica em outra premissa fundamental: as redes não seriam constituídas de nós igualitários, ou seja, com a possibilidade de ter mais ou menos o mesmo número de conexões. Ao contrário, tais redes possuiriam poucos nós altamente conectados (*hubs* ou conectores) em meio a uma grande maioria de nós com poucas conexões. Os *hubs* seriam os ricos, que tenderiam a receber sempre mais conexões. Redes com essas características foram denominadas por ele de "redes livres de escalas".

Segundo Wattenberg (2006), a análise de redes sociais envolve três tarefas fundamentais: (1) identificar comunidades: os atores devem ser agrupados em comunidades, de acordo com seus atributos. É importante avaliar a densidade de uma comunidade em termos de conexão e identificar cliques e relacionamentos abertos; (2) identificar atores centrais: é necessário identificar os atores que possuem maior número de conexões, assim como pontos de articulação – atores que formam pontes entre comunidades. Esta tarefa requer a compreensão da estrutura global da rede, isto é, encontrar comunidades, descobrir como são conectadas e quais atores as conectam entre si; (3) analisar papel e posições de conexões e atores: esta análise é realizada sabendo a conexão dos atores dentro e fora de uma comunidade. Essa tarefa requer interpretação e depende dos atributos de atores e relacionamentos.

3. Análise de Dados em Redes Sociais

Esta seção descreve uma proposta de processo para análise e visualização de dados de redes sociais, ilustrada na Figura 1, sendo uma versão simplificada aplicada a casos reais na Seção 4. Inicialmente, a estrutura de dados orientada para a representação de grafos e representando uma rede social é populada com dados brutos originados da aplicação em estudo. Como a proposta trata grandes volumes de dados, por questões de otimização do processo de aquisição de dados, estes dados devem ser armazenados para processamento futuro. Uma vez populada a estrutura de dados, uma representação do grafo pode ser visualizada de forma gráfica e interativa. Além disso, técnicas de descoberta de conhecimento e aprendizagem de máquina são utilizadas para inferir conhecimento implícito dos dados do grafo bem como predizer comportamentos futuros, respondendo questões do tipo "*se-o que-quando*" caracterizando a análise como uma aplicação de Inteligência Competitiva.

Por exemplo, considere uma rede social sobre produção científica de pesquisadores. Pode-se extrair informações estáticas de uma rede social, como, "*Qual o nível de colaboração de um determinado pesquisador?*", "*Qual o pesquisador mais produtivo no ano X?*" e informações temporais, como "*Qual a trajetória e tendência de carreira para um pesquisador com base em seu currículo atual?*". Informações estáticas podem ser inferidas com base na análise do grafo gerado a partir dos dados brutos, enquanto informações temporais podem ser obtidas através de descoberta de conhecimento, aperfeiçoadas com aprendizagem de máquina e através da interação direta com o grafo.

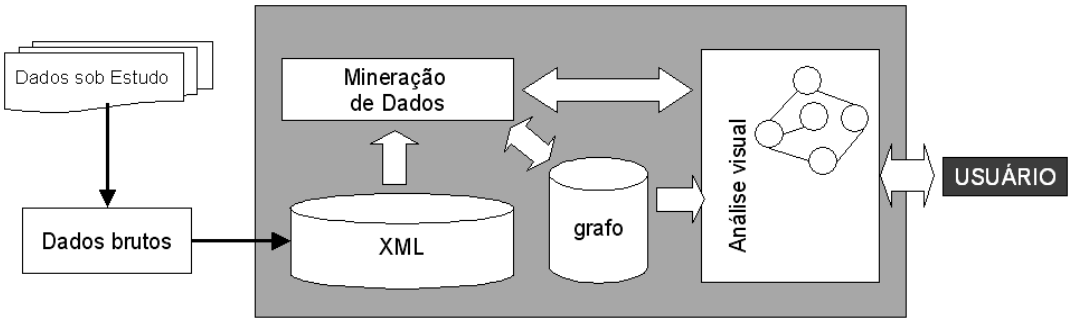

Figura 1. Análise e visualização de dados de redes sociais para extração de conhecimento.

O processo tratado neste artigo é composto por quatro componentes principais: (1) Gerenciamento de Dados – especificação de uma estrutura de dados adequada para gerenciar os dados da rede social em memória e armazenamento em meio físico; (2) Descoberta de Conhecimento – especificação de técnicas de mineração de dados, como predição e descrição, a fim de descobrir conhecimento implícito e futuro das redes sociais; (3) Aprendizagem de Máquina – em conjunto com as técnicas de descoberta de conhecimento, este componente atua racionalmente no ambiente com o objetivo de incrementar o desempenho em tarefas futuras; (4) Técnicas de Visualização – técnicas de visualização de informações que permitem a análise visual e interativa das redes sociais apoiadas pelos (e contribuindo para) processos de descoberta de conhecimento e aprendizagem de máquina.

3.1. Gerenciamento de Dados

A linguagem XML é amplamente aceita como o padrão para aplicações que envolvem troca de dados. Ao contrário de dados relacionais, dados XML não possuem esquema e são auto-descritivos, ao mesmo tempo que exibem uma estrutura hierárquica definida por elementos e atributos. Tal flexibilidade permite que XML seja empregada no desenvolvimento de aplicações em múltiplos contextos. Entre eles, pode-se citar a definição de padrões abertos para formato de arquivos, facilitando a troca de dados entre múltiplas aplicações em diferentes plataformas. Por exemplo, padrões de indústria especificados em XML incluem ACORD (seguros), ARTS (vendas), FpML (finanças), HL7 (saúde) e MISMO (empréstimos). Seja a base de dados localmente armazenada ou distribuída pela rede, seja estática ou adquirida através de *streams*, seja o processamento relacional ou nativo, as tarefas de armazenamento de dados e avaliação de consultas ainda são fatores fundamentais na otimização de qualquer sistema que utilize XML.

O armazenamento de dados e a avaliação de consultas XML apresentam vários desafios em relação ao processamento de dados relacionais. Primeiramente, dados XML possuem estruturas mais complexas em função da organização hierárquica intrínseca dos documentos XML. Assim, os dados seguem uma organização no formato de árvore, onde tarefas básicas (tais como indexação e consulta) que são realizadas eficientemente em estruturas planas (como listas e vetores) se tornam mais complexas e demoradas. Adicionalmente, dados XML não requerem um esquema rígido e fixo, tornando a organização dos dados ainda mais difícil. Finalmente, documentos XML são textuais e descritivos, nos quais a estrutura dos dados é definida para cada parte da informação,

resultando na repetição de identificadores XML. Tal repetição contribui para aumentar o tamanho das coleções de dados quando especificadas como um documento XML.

Considerando o armazenamento de dados XML, a solução inicial e mais simples é armazenar dados XML como CLOBs (*character large objects*) na base relacional. Nesse caso, cada documento XML fica armazenado como um todo. Porém, armazenar dados como CLOBs dificulta a tarefa de consulta sobre os dados (cada consulta requer um novo *parsing* do documento), dificulta a atualização parcial e impossibilita o controle da integridade dos dados. Outra solução, adotada na maioria dos sistemas gerenciadores de dados relacionais comerciais, é transformar dados XML em tabelas relacionais, processo conhecido como *shredding* [Florescu 1999]. Entretanto, *shredding* não gerencia modificações no esquema XML eficientemente e gera um *overhead* considerável no processamento de consultas devido às junções de várias tabelas. Desta forma, SGBDs XML nativos (com funções de armazenagem e consulta otimizadas para gerenciar dados XML) têm sido desenvolvidos [Jagadish 2002].

Numa rede social, os dados associados a nodos e arestas podem ser heterogêneos e semi-estruturados, traduzindo múltiplas relações. XML permite uma representação direta e fiel dos relacionamentos entre os objetos, facilitando o processo de visualização e descoberta de conhecimento. Além disso, o processo de aprendizagem e descoberta de conhecimento em redes sociais requer, em geral, a habilidade de explorar propriedades intrínsecas dos objetos e informações existentes nos (ou derivadas dos) relacionamentos entre os objetos da rede social. Neste contexto, a estrutura hierárquica do XML facilita o processo de mineração de dados. São desafios para a representação de XML.

- Representações lógicas e representações estatísticas – dois tipos de dependências podem ser identificadas no grafo que representa a rede social: dependência de ligações e dependência probabilística. O primeiro tipo representa o relacionamento lógico entre os diversos objetos das redes sociais, cujos fatores podem estar armazenados nos nodos ou nos relacionamentos entre os nodos. O segundo tipo representa os relacionamentos estatísticos como, por exemplo, a relação probabilística entre os atributos de objetos que estão relacionados na rede social. O gerenciamento dessas dependências é um desafio tanto para a especificação da estrutura de dados que representa a rede social quanto para o processo de extração de conhecimento desses dados;
- Construção do grafo – a construção do grafo deve considerar não somente a relação existente entre os objetos da rede social, mas também o relacionamento entre os atributos dos objetos relacionados na rede e os relacionamentos que possuem atributos associados. O desafio consiste em especificar uma estrutura simples, que contemple agregações, conjuntos, entre outras funcionalidades cuja representação facilite a manipulação e visualização;
- Representação de instâncias e classes – a estrutura de dados deve permitir a representação tanto de objetos individuais quanto de classes ou categorias de objetos;
- Uso de dados rotulados e não-rotulados – redes sociais devem incorporar tanto dados rotulados quanto não-rotulados. Um dado não-rotulado pode ajudar no processo de inferência da distribuição dos atributos dos objetos. As ligações entre os dados não-rotulados permitem utilizar atributos de objetos relacionados,

ao passo que ligações entre dados rotulados induzem dependências que podem auxiliar em inferências mais precisas.

3.2. Análise de Dados

Uma das etapas que se destaca no processo de descoberta de conhecimento é a mineração de dados – o processo de análise de conjuntos de dados que tem por objetivo a descoberta de padrões interessantes e que possam representar informações úteis. Um padrão pode ser definido como sendo uma afirmação sobre uma distribuição probabilística e pode ser expresso na forma de regras, fórmulas e funções, dentre outras. Como as redes sociais são dinâmicas e novos relacionamentos entre os objetos podem surgir a qualquer momento, técnicas de mineração de dados podem ser utilizadas para descobrir padrões existentes entre os objetos da rede social como também inferir futuros relacionamentos entre os objetos.

Duas técnicas fundamentais [Han e Kamber 2006] podem ser utilizadas para descoberta de conhecimento em redes sociais: descrição e predição. Os padrões descritivos são classificados em agrupamento, regras de associação e padrões seqüenciais, sendo utilizados, por exemplo, para encontrar padrões que sejam interpretáveis pelo homem e que descrevam os dados [John 1997]. Os padrões preditivos são definidos por regressão e classificação [Fayyad et al. 1996] e são utilizados para predizer o valor desconhecido ou futuro de um ou mais atributos com base no valor conhecido dos demais atributos. Métodos de descrição para redes sociais utilizam as técnicas convencionais de mineração de dados em bancos de dados. Métodos para predição buscam analisar a proximidade entre os objetos da rede, através da utilização conjunta de algoritmos de regressão e classificação e de técnicas da teoria de grafos.

As técnicas de descrição e predição [Han e Kamber 2006] são utilizadas em conjunto para analisar as seguintes evidências presentes em redes sociais:

- Classificação baseada em relacionamentos – nos métodos tradicionais de classificação, os objetos são classificados com base em seus atributos. Nas redes sociais, a classificação baseada no relacionamento existente entre os objetos pode ser utilizada para predizer a categoria de um objeto com base não somente nos seus atributos, mas também no relacionamento existente entre os objetos e entre os atributos dos objetos. No domínio de publicação científica, por exemplo, a predição pode inferir o tópico do artigo baseado tanto em palavras-chave quanto nos seus relacionamentos de citação;
- Predição de tipos de objetos – essa predição é baseada nos atributos e relacionamentos de um objeto e nos atributos dos objetos a ele relacionados. No domínio de publicação científica, por exemplo, uma tarefa pode ser definir se o tipo da publicação é um periódico, uma conferência ou um workshop;
- Predição de tipo de ligação – é o processo de descobrir o objetivo de um relacionamento baseado em suas propriedades. Por exemplo, pode-se querer descobrir se um relacionamento de orientador-orientando tem dois co-autores;
- Duplicação de objetos – é a tarefa de predizer se dois objetos são de fato o mesmo objeto, baseado em seus atributos e relacionamentos. Exemplos incluem

predizer se duas páginas Web são uma espelho da outra; se duas citações estão fazendo referência ao mesmo artigo, etc.;

- Detecção de grupos – é o processo de agrupamento, no qual pretende-se descobrir se um conjunto de objetos pertence ao mesmo grupo, baseado nos seus atributos e relacionamentos. Por exemplo, pode-se descobrir se um conjunto de pesquisadores publica em colaboração ou, mesmo que sem co-autoria, publicam sobre o mesmo tema;
- Detecção de subgrafos – busca por características em subgrafos dentro da rede social. Considerando uma rede de pesquisadores onde os relacionamentos são publicações científicas em co-autoria, a detecção de sub-grafos permite identificar com rapidez pesquisadores atuando em estreita colaboração.

3.3. Visualização

Genericamente, visualização é entendida como a "representação gráfica de dados ou conceitos" [Ware 2001], mas a aplicação desse termo é hoje associada à possibilidade de explorar as informações subjacentes à representação gráfica. Visualização de redes sociais é uma sub-área de visualização de informações, onde o conjunto de dados a ser visualizado e explorado é um grafo, com os nodos representando entidades sociais (pessoas, instituições, grupos, etc) e arestas representando os relacionamentos existentes. Para a visualização, portanto, o grafo deve ter sido previamente extraído da base de dados da aplicação, conforme discutido nas seções anteriores.

A visualização de grafos tem sido investigada há muitos anos e o desenho dessas estruturas é um problema clássico [Battista 1999]. Embora o problema possa ser resumido a "dado um grafo, encontrar posições para seus nodos e desenhar curvas conectando-os", esse não é um problema trivial, existindo uma vasta literatura sobre o assunto. Textos introdutórios, em português, podem ser encontrados em [Nascimento 2005] e [Freitas 2007]. Herman et al. (2000) apresentam uma excelente revisão sobre visualização de grafos, abordando tanto aspectos de leiaute de grafos como de interação e agrupamento. Um texto relativamente recente, de autoria de van Ham (2005), descreve a implementação de quatro técnicas de visualização de grafos, as quais são comparadas quanto à flexibilidade e escalabilidade.

No que se refere a leiaute, tanto as formas de diagrama de nodos e arestas bidimensional (2D) e tridimensional (3D) como as baseadas em matriz de adjacência têm sido utilizadas. As formas diagramáticas são providas por diversos algoritmos descritos na literatura [Battista 1999] e, de acordo com suas características, apresentam ótimo resultado para grafos esparsos. Porém, seu desempenho visual é bastante limitado para grafos densos. Assim, torna-se difícil para o usuário visualizar e interagir com os elementos do grafo.

A exploração da forma de matriz de adjacência por Henry e Fekete (2006) é suportada pela proposição de Bertin (1983) em usar essa representação para redes. Grafos baseados em matrizes de adjacência eliminam problemas de oclusão de vértices do diagrama, embora não sejam intuitivos para a maioria dos usuários. Apesar de ser eficaz na visualização de grafos densos, matrizes apresentam dificuldades quando se deseja percorrer caminhos nas redes sociais. Portanto, representações em matrizes são recomendadas para análise de comunidades, porém não são aconselháveis para a

visualização de estruturas globais. Henry et Fekete (2007) e Elmqvist et al. (2008) propuseram a combinação da forma de matrizes com uma forma diagramática, na tendência de técnicas híbridas [Sindre 1993] e de múltiplas visualizações coordenadas que se observa em muitos sistemas de visualização recentes.

Um dos atributos ainda pouco explorados, mas que possui grande potencial é a variação de informações em relação ao tempo. Em geral, redes sociais caracterizam-se pela quantidade de relacionamentos de uma pessoa. Porém, não se sabe a intensidade com que essas relações foram construídas. Se a visualização final retrata o total de relacionamentos de uma entidade ao longo de um período, não é possível atribuir o seu nível de atividade social. Ou, se a rede apresenta apenas informações atuais, perde-se o conhecimento em relação a eventos passados. Portanto, fatores externos que podem influenciar na intensidade de criação e rompimento de relações, como aspectos geográficos, acabam sendo mascarados.

Nesse contexto, é possível perceber a relação existente na aplicação em redes de co-autoria científica. Pajek [Nooy 2005] é uma ferramenta bastante utilizada para visualização de grandes redes. A principal característica é permitir a visualização de modo recursivo, decompondo a rede em estruturas menores e oferecendo ferramentas de análise de estruturas. Uma de suas características é a possibilidade de construir redes com atributos temporais. Kumar et al. (2006a) realizaram um estudo na área de visualização temporal de grafos. Utilizando dados da bolsa de valores dos Estados Unidos no período de 1990 a 2005, construíram uma ferramenta que permite a visualização de informações que variam ao longo do tempo. Nessa abordagem, a rede é subdividida em grupos de acordo com as áreas do mercado norte-americano. Esses grupos são conectados por arestas que representam o relacionamento escolhido. A navegação ao longo do tempo é realizada através de um cursor, onde o usuário pode explorar a rede manualmente, ou simplesmente através da execução de uma animação, mostrando assim a evolução temporal do grafo.

Aplicações complexas de visualização, como é o caso da análise visual de grafos densos, envolve a exploração de espaços multidimensionais de atributos, que é traduzida para uma gama de tarefas em diferentes níveis de abstração [Zhou 1998, Amar 2005], dentre as quais as tarefas analíticas tem papel importante. Nesse sentido, a integração do resultado das análises à visualização é um aspecto essencial de ferramentas de análise de redes de sociais. Vários caminhos de navegação, alternativas de filtragem e derivação de medidas são possíveis. A exploração do espaço multidimensional, em geral desdobra-se em várias tarefas diferentes como identificação, localização, comparação, computações diversas, o que introduz dificuldades na escolha das técnicas de interação mais apropriadas para dar suporte a essas tarefas variadas. Além disso, para que a interação seja ao mesmo tempo eficiente e confortável, depende-se de uma boa combinação de dispositivo de entrada, tarefa interativa a ser executada e dispositivo e técnica de saída. Técnicas de interação com multimodalidade vêm sendo consideradas como uma maneira de alcançar considerável aumento na satisfação e conforto do usuário através de uma interação natural. A exploração de vários canais de comunicação entre um sistema e seus usuários pode ter um grande impacto na sua eficiência de utilização em aplicações onde o conjunto de dados é volumoso e/ou complexo.

4. Estudo de Caso

Para discutir as idéias mencionadas, foi conduzido um estudo de caso com base na análise da relação de co-autorias entre pesquisadores, medida através da publicação de artigos científicos. No contexto mais amplo, está em análise a produção científica de pesquisadores e grupos. Esse é um tema relevante no meio científico tal que melhorias nas formas de apresentação e, principalmente, de análise de dados da produção científica certamente contribuirão para o crescimento da produção nacional pelo entendimento a respeito do processo de produção qualificada que os resultados advindos de novas formas de análise podem trazer.

O estudo de caso foi conduzido com base numa ferramenta de visualização que permite consultas através da interação e manipulação do grafo [Spritzer 2008]. Os dados foram extraídos da BDBCOMP e DBLP e representados em XML: dois conjuntos contendo dados das publicações das duas bases de dados e um conjunto contendo dados das publicações na BDBCOMP por docentes do Instituto de Informática da UFRGS. Como o interesse no estudo de caso é uma visão das co-autorias e sua influência na produtividade dos pesquisadores, o grafo foi construído tendo como nodos os pesquisadores e como arestas as colaborações. Como atributos do nodo estão o nome do pesquisador e o total de artigos registrados na base acessada; na aresta, os atributos são os identificadores dos artigos (e os anos de publicações dos mesmos) produzidos em conjunto pela dupla de pesquisadores relacionados.

Foram utilizados 3 grafos, cujas visualizações iniciais são mostradas na Figura 2. As visualizações foram geradas com a ferramenta de Spritzer e Freitas (2008) usando uma adaptação do algoritmo de Fruchterman e Reingold (1991), baseado em forças que atuam até que o leiaute atinja uma situação de equilíbrio dinâmico, posicional. Pode-se observar alguns poucos nodos bem marcados, que correspondem aos pesquisadores com maior número de colaboradores (nodos com grau elevado). Mesmo não sendo grafos muito grandes, a profusão de arestas torna impossível visualizar claramente os nodos correspondentes a esses pesquisadores. A ferramenta utilizada para visualização oferece a possibilidade de separar os nodos que atendam critérios. Isso é feito criando “imãs” cuja força de atração de nodos está relacionada com uma expressão de busca. No caso do presente exemplo, essa facilidade foi utilizada para responder uma pergunta básica: “Quais pesquisadores têm maior número de colaboradores?” utilizando o cálculo do grau dos nodos como requisito para separar esses pesquisadores numa região do grafo. A resposta utilizando o grafo que representa os congressos no DBLP pode ser vista na Figura 3a e a mesma questão para a BDBCOMP é visualizada na Figura 3b.

Considerando que os atributos de uma aresta são os artigos que dois pesquisadores tem em comum, com o ano da publicação associado a cada artigo pode-se avaliar a evolução temporal da colaboração entre esses pesquisadores. Outra análise interessante seria determinar se existe relação entre produtividade e quantidade de colaboradores, ou seja, os pesquisadores que mais colaboram são também os que mais publicam? Isso pode ser feito, na ferramenta utilizada, com dois imãs, um para atrair nodos com grau acima de um determinado valor, e outro para atrair nodos com atributo “número de artigos” acima de um determinado valor. Os nodos que atendessem os dois requisitos ficariam na região entre os imãs.

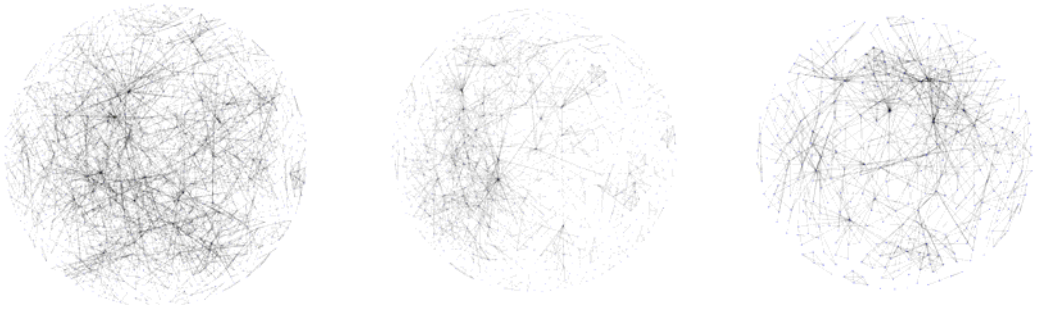

Figura 2. Visualizações dos grafos utilizados no estudo de caso: co-autorias em congressos indexados no BDBCOMP (930 nodos, 2043 arestas), co-autorias em congressos indexados na DBLP (675 nodos, 1395 arestas) e co-autorias em congressos indexados no BDBCOMP com docentes do Instituto de Informática (357 nodos e 853 arestas).

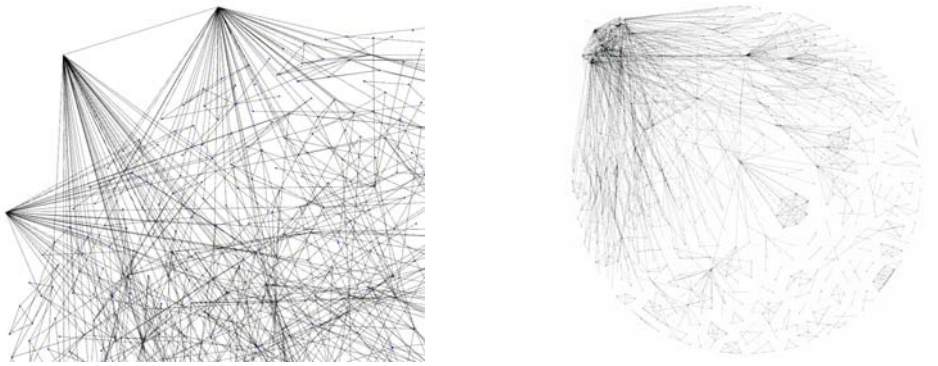

Figura 3. Visualização de parte do grafo com os pesquisadores que possuem mais de 30 colaboradores (base de dados BDBCOMP) e o pesquisador com mais de 48 colaboradores na base de dados DBLP (3 pesquisadores apenas tem mais de 45 colaboradores). Exemplos adicionais: http://www.inf.ufrgs.br/cg/gallery.html

Tendo como atributo dos nodos a instituição dos pesquisadores, outra possibilidade de análise seria observar a característica de um pesquisador em relação à colaboração. Quem são os pesquisadores que têm mais colaboração interna (com pesquisadores da mesma instituição)? E externa? Quem é mais produtivo, quem colabora internamente ou externamente? Quem mais colabora internamente é também quem tem mais colaboração externa? Outras questões certamente exigirão mais atributos e técnicas mais elaboradas de mineração de dados. Qual o perfil de um pesquisador produtivo analisando a sua produção ao longo do tempo? Um grupo é produtivo porque tem vários pesquisadores medianos ou por que tem alguns poucos muito produtivos? Qual a visibilidade da qualidade da produção em uma rede? A qualidade é mensurável de forma absoluta ou relativa?

Este estudo de caso levanta apenas umas poucas questões de um cenário mais amplo de análise de co-autorias. Os resultados apresentados nesta seção são fruto de um trabalho em andamento. Atualmente, estão sendo investigadas alternativas para a visualização e análise de dados temporais, a fim de obter subsídios para responder às perguntas colocadas acima.

5. Cenários e Desafios em Análise de Redes Sociais

No contexto de redes sociais naturais, uma vez que existe um grande interesse na análise de propriedades e da interação entre redes de agentes, é natural que se estude os mecanismos de aprendizagem coletiva e individual em redes sociais. Entre os estudos recentes, incluem-se a análise do papel das propriedades de agentes no desempenho global da rede em tarefas de aprendizagem, como os agentes aprendem através da interação, como solucionam problemas ou como otimizam tarefas [Shinoda 2007, Heck 2007, Araujo e Lamb 2007].

No caso de redes de cooperação acadêmica, mecanismos de aprendizagem de máquina poderiam permitir a identificação de comunidades de agentes (pesquisadores), o acompanhamento da propagação de informações intra/entre-grupos (incluindo como agentes se comunicam e comunicam os resultados de pesquisa entre eles, como constroem cooperações e projetos), como são agrupadas as tarefas (projetos, linhas de pesquisa, financiamento de pesquisa, medidas de eficiência em projetos, medidas de produtividade, entre outras). Todas estas análises podem ser direcionadas tanto a grupos de agentes (sub-redes), quanto a indivíduos. Além disso, as medidas de tráfego de informações, a quantidade de interações e a complexidade dos mecanismos de interação coletiva e individual também se apresentam como desafios a serem vencidos para que mecanismos de aprendizagem em redes sociais tornem-se computacionalmente úteis nos cenários aqui apresentados.

Uma tendência emergente, adotada por empresas como a Amazon, por exemplo, é a utilização de grupos de pessoas para prestar serviços online de forma distribuída, como classificação e organização de dados [Economist 2006]. Quando permite-se que tais pessoas interajam, formam-se redes sociais, cujas propriedades podem influenciar a capacidade do grupo em solucionar problemas de forma eficaz. É desejável estudar o comportamento de redes sociais que visam resolver problemas de forma distribuída. Modelos baseados em inteligência artificial permitem facilitar, através de simulações numéricas, a análise de tais redes. Estes modelos necessariamente envolvem áreas como sistemas multi-agentes e aprendizagem de máquina, de forma a permitir a modelagem de grupos de atores cognitivos.

Por outro lado, algoritmos de aprendizado de máquina podem ser aplicados diretamente à análise de dados de redes sociais arbitrárias, auxiliando na identificação e na classificação automática de grupos e comunidades imersos em redes sociais [Roffilli e Lomi 2006]. Outra possibilidade é a de áreas de inteligência artificial, como a "computação inspirada em redes sociais", que se beneficiaram da análise de redes sociais para construir novos métodos para solucionar problemas diversos. É o caso de algoritmos de aprendizagem de máquina que fazem uso de múltiplos agentes que trocam informações através de um modelo de rede social [Araújo 2008]. Questões relevantes incluem determinar propriedades de redes sociais que têm efeito sobre o desempenho de tais sistemas, bem como a análise de recursos computacionais e sociais dispendidos na execução de tarefas.

6. Conclusões

O objetivo deste artigo é discutir requisitos e desafios em descoberta de conhecimento em redes sociais utilizando técnicas de mineração de dados, inteligência artificial,

aprendizagem de máquina e visualização interativa de informações. O cenário utilizado como estudo de caso desta proposta foi a aplicação de análise de redes sociais para visualização e avaliação da colaboração científica entre pesquisadores.

A tarefa de avaliar a produção científica de um pesquisador e grupos de pesquisadores é baseada fortemente na análise de currículo. É o que fazem, por exemplo, as agências de fomento à pesquisa e desenvolvimento ou comissões de avaliação, quando necessitam considerar a produção científica dos pesquisadores no processo de concessão de bolsas e auxílios, na seleção de consultores e membros de comitês, na aprovação de projetos ou ainda na avaliação do conceito de um programa de pós-graduação. Ferramentas visuais que permitam análise e descoberta de conhecimento que levem em consideração aspectos qualitativos e quantitavos podem oferecer subsídios na correta aplicação dos recursos oportunizados.

Um exemplo de aplicação da proposta aqui apresentada, seria a análise da plataforma Lattes. Apesar desta ser o mais importante instrumento de armazenamento dos dados referentes à produção científica de pesquisadores brasileiros, ela possui um mecanismo de representação de dados que não permite nem o cruzamento de dados entre pesquisadores, nem consultas visuais e interativas. O processo aqui proposto poderia oportunizar a análise preditiva do comportamento de pesquisadores, baseada no histórico de sua produção científica presente na base de dados Lattes. Acredita-se que, através de consultas interativas e visuais, pode-se analisar com mais facilidade a carreira de um pesquisador, propor aconselhamentos com base na análise comparativa de dados de pesquisadores bem sucedidos e avaliar a produção através de métodos mais confiáveis e eficazes.

Fundamentalmente, o processo proposto para solucionar os problemas relacionados à extração de informações de redes sociais remete ao grande desafio da gestão da informação em grandes volumes de dados, mais especificamente à recuperação e disseminação de informação relevante. Como foi apresentado em detalhe na Seção 3, entende-se que a busca de uma solução robusta para extração de informações em grandes bases de dados (neste caso redes sociais), só é possível através do trabalho combinado entre diferentes áreas de pesquisa.

Técnicas apropriadas para gerenciamento de dados e descoberta de conhecimento são fundamentais para permitir não só o armazenamento das informações de forma adequada como a extração de dados relevantes. Entretanto, a grande quantidade de informações tratada neste estudo de caso demonstra que o avanço das pesquisas nestas áreas em específico não é suficiente para que se atinja um resultado satisfatório. Propõe-se assim a exploração de técnicas interativas de visualização de informações não apenas como a etapa final no processo de recuperação de informações, mas eventualmente como uma etapa capaz de realimentar o processo de mineração de dados. Finalmente, vislumbrou-se o uso de mecanismos de aprendizagem de máquina para permitir análises complementares dos dados disponibilizados.

Os resultados preliminares apresentados neste artigo, bem como a proposta de processo de análise de redes sociais apresentada são o produto de um trabalho em equipe. Acredita-se que a maior contribuição está na heterogeneidade da solução apresentada.

Agradecimentos

Agradecemos ao CNPq o apoio financeiro, sob diversas formas, aos autores deste artigo.

Referências

Amar, R., Eagan, J. e Stasko, J. (2005) "Low-Level Components of Analytic Activity in Information Visualization". Proc. IEEE Symp. on Inform. Visualization, pp. 111-147.

Araújo, R. M., e Lamb, L. C. (2007) “An Information-Theoretic Analysis of Memory Bounds in a Distributed Resource Allocation Mechanism”. In Proc. of Intl. Joint Conf. on Artificial Intelligence IJCAI-07, pp. 212-217. AAAI Press.

Araújo, R. M., e Lamb, L. C. (2008) “Distributed Problem Solving by Memetic Networks”. In Proc. of GECCO 2008, to appear.

Barabasi, A.-L. (2003) “Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life”. Plume Books.

Battista, G., Eades, P., Tamassia, R. e Tollis, I.G. (1999) “Graph Drawing: Algorithms for the Visualization of Graphs”. New Jersey: Prentice Hall.

Bertin, J. (1983) “Semiology of Graphics”, University of Wisconsin Press.

Day, G. S. e Wensley, R. (1988) “Assessing advantage: a framework for diagnosing competitive superiority”. Journal of Marketing, p. 1-20, V.52, n.1.

DBLP: Digital bibliography and library project, http://dblp.uni-trier.de, 2007.

Degenne, A. e Forse, M. (1999) “Introducing Social Networks”. Sage Publications.

The Economist (2006) "Artificial artificial intelligence", London, 8 June.

Elmqvist, N., Henry, N., Richie, Y., Fekete, J-D (2008). “ Melange: Space folding for multi-focus interaction”. In ACM SIGCHI, New York. ACM.

Erdös, P. e Rényi, A. (1960). “On the evolution of random graphs”. In Publ. Math. Inst. Hungar. Acad. Sci., pages 17–61.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. (1996) “From Data Mining to Knowledge Discovery: An Overview”. In: Fayyad, U. M. et al. Advances in Knowledge Discovery and Data Mining. Menlo Park: AAAI Press. 611p. p.11-34.

Florescu, D. e Kossmann, D. (1999) “Storing and Querying XML Data using an RDBMS”. IEEE Data Engineering Bulletin, 22(3), 27-34.

Fruchterman, T.M.J. e Reingold, E.M. (1991). "Graph Drawing by Force–Directed Placement". Software - Practice & Experience, 21 (Nov), 1129–1164

Freitas, C.M.D.S. (2007) “Visualização de Informações e a Convergência de Técnicas de Computação Gráfica e Interação Humano-Computador”. In: Kowaltowski, T. e Breitman, K. (org.) Atualizações em Informática. Rio de Janeiro, PUC-Rio/SBC.

van Ham, F. (2005) “Interactive Visualization of Large Graphs”. PhD Thesis,Technische Universiteit Eindhoven, ISBN 90-386-0704-0

Han, J. e Kamber, M. (2006) “Data Mining: concepts and Techniques”. Morgan Kaufmann, Second Edition. 600p.

Henry, N., e Fekete, J.-D. (2006) “MatrixExplorer: a dual-representation system to explore social networks". IEEE Trans. on Visualization & Computer Graphics, 12(5): 677-684.

Henry, N. e Fekete, J.-D. (2007) “Nodetrix: Hybrid representation for analyzing social networks”. IEEE Transactions on Visualization & Computer Graphics, 13(6): 1302-1309.

Herman, I, Melançon, G. e Marshall, M.S. (2000) “Graph Visualization and Navigation in Information Visualization: A Survey”. IEEE Transactions on Visualization & Computer Graphics, 6(1):24-42.

Jagadish, H. V. et al. (2002). “TIMBER: A native XML database”. VLDB Journal, 11(4):274-291.

John, G. H. (1997) “Enhancements to the Data Mining Process”. Stanford Univ., Ph.D. Thesis.

Kumar, G. e Garland, M. (2006a). “Visual exploration of complex time-varying graphs”. IEEE Transactions on Visualization & Computer Graphics, 12(5):805–812.

Kumar, K., Novak, J., e Tomkins, A. (2006b) “Structure and evolution of online social networks”. In Proc. of ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, p. 611- 617, New York.

Laender, A. H. F.; Gonçalves, M. A.; Roberto, P. A. (2004) “BDBComp: building a digital library for the Brazilian computer science community”. In: 4th ACM/IEEE-CS Joint Conf. on Digital Libraries, New York, NY, USA. p. 23–24.

Nascimento, H.A.D e Ferreira, C.B.R. (2005) “Visualização de Informações – Uma Abordagem Prática”. In Anais do XXV Congresso da Soc. Bras. de Computação, pp. 1262-1311.

Nooy, W. de, Mrvar, A. e Batagelj, V. (2005) “Exploratory Social Network Analysis with Pajek” (Structural Analysis in the Social Sciences). Cambridge University Press.

Roffilli, M. e Lomi, A. (2006), "Identifying and classifying social groups: A machine learning approach", in Data Science and Classification, V.Batagelj, H.-H.Bock, A.Ferligoj, A.Ziberna (Editors). Springer.

Sindre, G., Gulla, B. e Jokstad, H. G. (1993). “Onion graphs: aesthetics and layout”. In VL, pages 287–291.

Shinoda, K., Matsuo, Y. e Nakashima, H. (2007). “Emergence of Global Network Property Based on Multi-Agent Voting Model”. In Proc. AAMAS2007, ACM Press.

Spritzer, A.S. e Freitas, C.M.D.S. (2008) "A Physics-based Approach for Interactive Manipulation of Graph Visualizations". Proc. Intl. Conf. on Advanced Visual Interfaces, AVI 2008, Napoli.

Ware, C. (2001) “Information Visualization: Perception for Design”, San Francisco, Morgan Kaufmann.

Wasserman, S. e Faust, K. (1994) “Social network analysis: methods and applications”, vol. 8 of Structural analysis in the social sciences. Cambridge Univ. Press, Cambridge.

Wattenberg, M. (2006) “Visual exploration of multivariate graphs”. In Proc. of SIGCHI Conf. Human Factors in Computing Systems, p. 811–819, New York. ACM.

Watts, D. J. (2003) “Six Degrees: The Science of a Connected Age”. W.W. Norton & Company.

Zhou, M. e Feiner, S. K. (1998) "Visual Task Characterization for Automated Visual Discourse Synthesis". Proc. CHI’98 Conference, ACM Press, pp. 392-399.