

Prospecção de dados acadêmicos de Currículos Lattes através de Scriptlattes

Jesús Pascual Mena-Chalco
Roberto Marcondes Cesar Junior

Introdução

A prospecção de dados é um processo de extração e exploração de grandes volumes de dados, geralmente utilizada para identificar ou evidenciar possíveis relacionamentos entre instâncias dos elementos tratados (YE, 2003). A extração de dados de produção científica, identificação de padrões bibliométricos, e modelagem e visualização efetiva de redes de interação entre coautores são tópicos relevantes na área de Bibliometria e Cientometria. Nos últimos anos, tem sido dado especial interesse a tais tópicos devido à descoberta de conhecimento que pode ser obtida a partir do tratamento de conjuntos de dados disponíveis nos repositórios de produção científica (e.g. banco de dados de produções bibliográficas, de orientação acadêmica, de projetos de pesquisa, e de diretórios de grupos de pesquisa).

Por outro lado, no Brasil, o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) realiza um importante trabalho na integração de bases de currículos acadêmicos de instituições públicas e privadas em uma única plataforma denominada Lattes. Os chamados "Currículos Lattes" são considerados um padrão nacional de avaliação, representando um histórico das atividades científicas, acadêmicas e profissionais de pesquisadores cadastrados na plataforma (AMORIN, 2003). Os currículos Lattes foram projetados para mostrar informação pública, individual, de cada usuário cadastrado na plataforma. Nesse contexto, muitas vezes, realizar uma compilação ou summarização de produções bibliográficas para um grupo de usuários cadastrados de médio ou grande porte (e.g. grupo de professores, departamento de pós-graduação) realmente requer um grande esforço manual suscetível a falhas. Assim, o scriptLattes (MENA-CHALCO; CESAR-JR, 2009), uma ferramenta de software livre, foi projetado para a extração e compilação automática de produções

bibliográficas, técnicas e artísticas, orientações, projetos de pesquisa, prêmios e títulos, além de possibilitar a geração de grafo de colaborações e mapa de geolocalização de um conjunto de pesquisadores cadastrados na plataforma Lattes.

O scriptLattes descarrega automaticamente os currículos Lattes (em formato HTML) de um grupo de pessoas de interesse, compila as listas de produções, tratando apropriadamente as produções duplicadas e similares. Em seguida, são gerados relatórios, em formato HTML, com listas de produções e orientações separadas por tipo e colocadas em ordem cronológica invertida. Adicionalmente, a ferramenta permite a criação automática de grafos (redes) de coautoria entre os membros do grupo e um mapa de geolocalização dos membros e alunos (de pós-doutorado, doutorado e mestrado) com orientação concluída. (SCRIPTLATTEs, 2011)

No nosso entendimento, essa ferramenta de software livre é a pioneira na prospecção de extensos conjuntos de dados acadêmicos provenientes de Currículos Lattes em formato HTML, e atualmente está sendo útil para extrair e representar conhecimento de grupos de pessoas cadastradas na plataforma Lattes, de forma simples. Esse conhecimento pode ser usado para explorar, identificar ou validar padrões de atividades científicas, trazendo assim informação bibliométrica e/ou cientométrica sobre um grupo de interesses (NICHOLSON, 2006) (PENG; MCCALLUM, 2006). A relevância deste trabalho recai sobre as vantagens decorrentes da utilização do processo considerado na ferramenta para a realização de análises consolidadas das produções científicas e das relações entre os atores da academia (KLINK et al., 2006) (KOUZES et al., 2009).

Descrição da ferramenta

Atualmente, o scriptLattes é um programa desenvolvido na linguagem de programação Python e está composto de seis módulos. Uma descrição detalhada de todos os módulos encontra-se em Mena-Chalco e Cesar-Jr. (2009). A Figura 1 mostra a interação entre os dados de entrada e saída, e das duas plataformas consideradas para a consulta de informações: Plataforma Lattes e Plataforma de geolocalização

(Google Maps). Nas seções seguintes descrevemos as principais características dos módulos projetados.

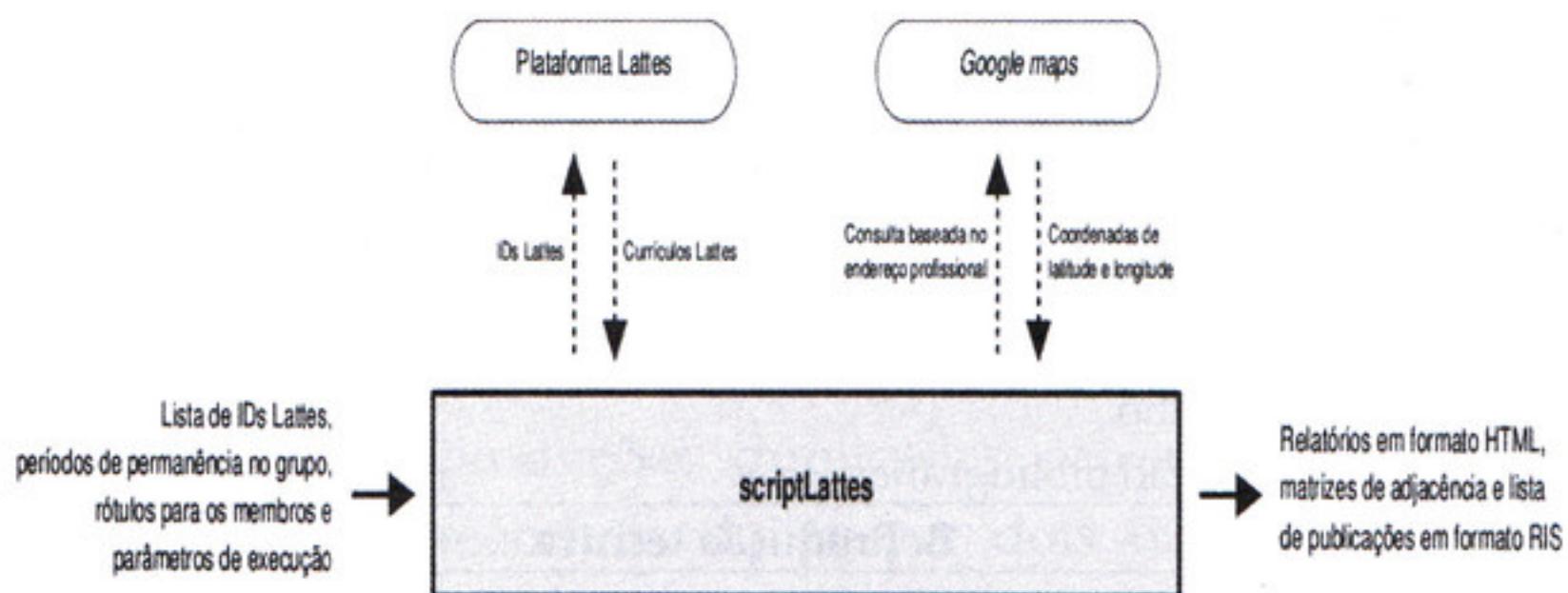


Figura 1 - Diagrama de fluxo de informações considerado no scriptLattes

Dados de entrada

A entrada para o programa está composta por uma lista ASCII de IDs de currículos Lattes (código de 16 dígitos que o CNPq utiliza como identificador de cada currículo Lattes), conjuntamente com o período de permanência no grupo, isto é, os anos em que cada membro foi associado ao grupo (e.g. grupo de pesquisa, departamento de pós-graduação), e um rótulo/etiqueta que é utilizado na visualização do grafo de colaborações (cada rótulo diferente é representado por uma cor diferente).

Através de um *parser* HTML (determinístico) - programa, baseado em análise textual, que permite identificar e extrair regiões ou trechos específicos de texto (TOMITA, 1991) - são automaticamente extraídos de cada currículo Lattes, indicado no arquivo de entrada para o programa, os dados correspondentes ao: nome completo do membro, nome em citações bibliográficas, endereço profissional, tipo de bolsa de produtividade, foto, sexo e data de atualização do currículo. Adicionalmente, são extraídas as listas completas de produções acadêmicas pertencentes ao período de permanência. (Quadro 1)

A. Produção bibliográfica
Artigos completos publicados em periódicos
Livros publicados/organizados ou edições
Capítulos de livros publicados
Textos em jornais de notícias/revistas
Trabalhos completos publicados em anais de congressos
Resumos expandidos publicados em anais de congressos
Resumos publicados em anais de congressos
Artigos aceitos para publicação
Apresentações de trabalho
Demais tipos de produção bibliográfica
B. Produção técnica
Softwares com registro de patente
Softwares sem registro de patente
Produtos tecnológicos
Processos ou técnicas
Trabalhos técnicos
Demais tipos de produção técnica
Total de produção técnica
C. Produção artística
D. Supervisões e orientações em andamento ou concluídas
Supervisão de pós-doutorado
Tese de doutorado
Dissertação de mestrado
Trabalho de conclusão de curso de graduação
Iniciação científica
Orientações de outra natureza
E. Projetos de pesquisa
F. Prêmios e títulos
G. Eventos (participação e organização)

Quadro 1 - Tipos de produção acadêmica extraídas dos currículos Lattes

É importante destacar que um desafio computacional para o programa é o tratamento dos dados em formato HTML, onde as partes constituintes das produções acadêmicas (e.g. nomes dos autores, título da publicação, título do projeto, nome do meio da publicação, número de páginas, volume, páginas, ano) são apresentadas sem alguma indicação de separação. Assim, o *parser* desenvolvido identifica, na grande maioria dos casos, todas as partes constituintes das produções acadêmicas. É relevante destacar ainda que as listas de todas as produções acadêmicas descritas no Quadro 1 são limitadas pelo período de permanência.

Tratamento de redundâncias

Várias produções acadêmicas são frequentemente elaboradas em colaboração com um ou mais pesquisadores do mesmo grupo. Uma produção (e.g. artigo completo publicado em periódico) pode aparecer duplicada nos relatórios, dado que ambos colaboradores são coautores. O programa desenvolvido mantém um módulo de tratamento de redundâncias que permite a detecção de produções acadêmicas iguais ou similares. Assim, as produções duplicadas são usadas para detectar colaboração entre os membros do grupo: dois ou mais membros são considerados como colaboradores se existe uma produção comum entre eles.

A detecção de produções similares é realizada através de comparações dois a dois entre todas as produções de conjuntos de dados separados por ano e tipo de produção (por ex., artigo publicado em periódico ou capítulo de livro), de tal forma que produções com anos de publicação diferentes não sejam utilizadas em nenhuma comparação, permitindo assim uma diminuição substancial de tempo de processamento do módulo de tratamento de redundâncias.

Devido a inconsistências como erros de digitação ou falta de padronização na escrita dos nomes dos coautores (KANG et al., 2009) no preenchimento das informações nos currículos Lattes, a comparação de duas produções quaisquer é realizada através de um casamento aproximado entre os títulos associados a cada cadastro. Atualmente, duas publicações são consideradas iguais se a porcentagem de similaridade entre os títulos for maior a uma determinada porcentagem. A similaridade entre duas cadeias baseia-se na distância proposta por Levenshtein (NAVARRO, 2001). A distância Levenshtein é obtida através do número mínimo de inserções, eliminações ou substituições de caracteres necessários para transformar um texto em outro, sendo que a distância Levenshtein zero (0) indicará que dois títulos analisados são exatamente iguais.

Para nossos testes, consideramos dois títulos equivalentes se ambos são pelo menos 80% similares. Esse valor pode ser facilmente configurado no programa para limitar a porcentagem de similaridade entre as produções acadêmicas.

Uma das características importantes desse módulo é a conjunção de dados nas produções iguais ou similares, de tal forma que as informações faltantes de um cadastro possam ser combinadas e/ou complementadas com as informações do outro. Atualmente, a complementação refere-se apenas à utilização exclusiva do campo com maior tamanho ou comprimento textual. Com este módulo, o scriptLattes é capaz de manter um registro de todos os coautores (pertences ao grupo em análise) associados a uma determinada produção acadêmica. Note que esta informação será importante na ponderação numérica de produções acadêmicas, como por exemplo, a normalização dos pesos nas arestas dos grafos de colaboração.

Grafos de colaboração

Geralmente, um grafo de colaborações/coautoria mostra atividades acadêmicas que são realizadas de forma conjunta por membros de um grupo (KLINK et al., 2006; MAIA; CAREGNATO, 2008). O programa desenvolvido usa um grafo (ou rede) para representar a colaboração entre membros de um grupo baseados exclusivamente na sua produção bibliográfica, técnica ou artística (orientações acadêmicas, prêmios e/ou títulos, e projetos de pesquisa não são considerados nos grafos de colaboração).

É válido destacar que nos últimos anos, as características de: (1) alto coeficiente de *clusterização*, (2) comportamento de um mundo pequeno e (3) distribuição *scale-free* foram associadas aos grafos de colaborações próprias de redes sociais (BARABASI; ALBERT, 1999). Cada membro é representado por um nó se e somente se uma produção acadêmica em comum dos membros é detectada como produção redundante no módulo de tratamento de redundâncias.

Nos relatórios gerados, são mostrados três tipos de grafos referentes a: (i) grafos de colaboração (não direcionado) sem pesos, em que as arestas representam apenas as ligações de trabalho colaborativo; (ii) grafos de colaboração (não direcionado) com pesos, em que o peso de uma aresta representa o número de produções acadêmicas elaboradas em coautoria entre dois nós, e (iii) grafos de colaboração (direcionado) com pesos normalizados, em que os pesos das arestas salientes de um dado nó (membro) são normalizados pela quantidade

total de produções acadêmicas feitas em colaboração, como sugerido por Liu et al. (2005). Na Figura 2 é possível visualizar um exemplo de grafos de colaborações criados a partir de três publicações elaboradas por quatro autores.

Produções elaboradas em colaboração

Artigo 1: Elaborado pelos autores M1 e M2.
 Artigo 2: Elaborado pelos autores M1, M2 e M3.
 Artigo 3: Elaborado pelos autores M1 e M4.

Autor M1: Participa em 3 artigos.
 Autor M2: Participa em 2 artigos.
 Autor M3: Participa em 1 artigo.
 Autor M4: Participa em 1 artigo.

Grafos de colaborações

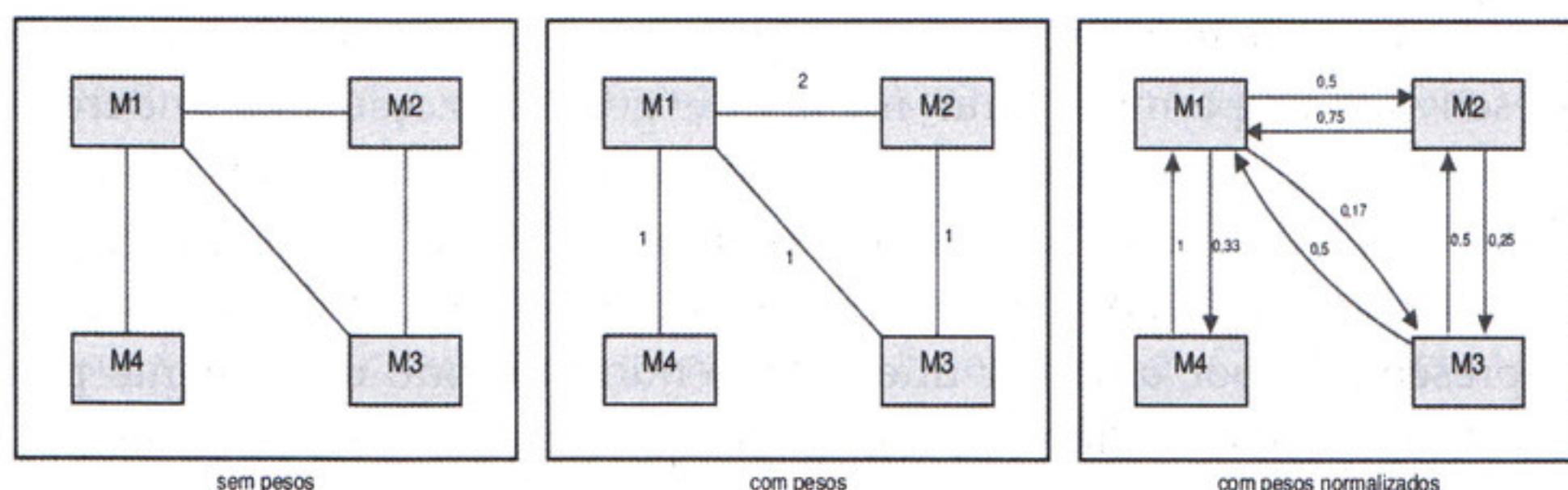


Figura 2 - Exemplo de grafos de colaboração criados a partir da detecção de 3 artigos elaborados por 4 autores (membros): M1, M2, M3 e M4

A normalização permite atribuir maior peso para autores que coproduziram mais publicações em conjunto. Os pesos normalizados intuitivamente dão uma ideia da 'importância' de um coautor na produção realizada em colaboração com outro. Por exemplo, para M2 o colaborador M1 participa em 75% da sua produção feita em colaboração, isto é, M2 é 75% importante para M1. Já para M1, M2 é apenas 50% importante para sua produção feita em colaboração. Por outro lado, para M4, M1 é 100% importante na sua produção (colaborativa) acadêmica, entretanto para M1, M4 é importante apenas 33%. Note que este último comportamento é típico nas relações orientador-orientado (M1–M4).

Os pesos do grafo direcionado, além de mostrar a importância de colaboração e sua reciprocidade, também são utilizados no scriptLattes como base para o cálculo dos graus de colaboração (LIU et al., 2005) dos membros do grupo. Entenda-se como grau de colaboração, um valor numérico que representa o impacto de um membro no grafo de colaborações (o algoritmo proposto por Liu et al.

(2005), denominado *AuthorRank*. Trata-se de uma adaptação do algoritmo *PageRank* utilizado no sistema de busca de páginas relevantes no buscador Google). Dessa forma, os membros de maior impacto colaborativo no próprio grupo terão os maiores graus de colaboração, isto é, quanto maior grau de colaboração, mais participativo o membro será no grupo em análise.

Mapa de geolocalização

Frequentemente, é desejável conhecer a localização geográfica atual dos membros de um grupo. Nesse contexto, o programa desenvolvido permite gerar mapas de geolocalização dos endereços profissionais tanto dos membros do grupo, quanto dos alunos formados pelo grupo, isto é, alunos com orientação concluída de pós-doutorado, doutorado e mestrado. No mapa, cada tipo de orientação é representado por uma cor diferente, sendo utilizado comumente para mostrar a influência/impacto do grupo na formação de profissionais. O endereço profissional de um aluno orientado é extraído desde que o orientador (membro do grupo) tenha cadastrado o ID Lattes do aluno no próprio currículo Lattes.

A plataforma *on-line* do *Google Maps* é utilizada para obter, de forma automática, as coordenadas de geolocalização em termos de latitude e longitude, considerando como parâmetros de consulta o CEP, UF e o nome do país. Veja em Enkhsaikhan; Liu; Reynolds (2008) uma abordagem similar, utilizada na visualização geográfica e temporal para grafos de coautoria. Uma geolocalização de uma pessoa não terá representação no mapa caso o endereço tratado seja incorreto ou este não estiver cadastrado na plataforma Lattes.

No Brasil, em casos específicos a Empresa Brasileira de Correios e Telégrafos define CEPs especiais que a plataforma do *Google Maps* não os interpreta corretamente. Nesse sentido, no scriptLattes é definido um procedimento que permite trocar, através de um dicionário, CEPs especiais por CEPs especificados no *Google Maps*. Esta correção de CEPs ajuda a refinar a localização geográfica para alguns endereços profissionais.

Geração de relatórios

A saída do sistema é um conjunto de relatórios, em formato HTML, referentes à compilação de dados em termos de produção científica. O formato HTML foi escolhido para todos os relatórios por ser um formato padrão para visualização de informação na internet.

Os relatórios são separados por tipos e mostram uma informação quantitativa classificada por ano em ordem cronológica invertida correspondente a: (i) Produções bibliográficas, técnicas e artísticas, (ii) Orientações em andamento e concluídas, (iii) Projetos de pesquisa, e (iv) Prêmios e títulos. Todos os relatórios mostram um gráfico de barras com o número de produções discretizados por ano. Os tamanhos das barras são proporcionais aos valores de produção acadêmica do grupo. Também são mostrados, para cada produção acadêmica, *links* diretos para buscas em alguns dos principais buscadores de citações disponíveis na internet (e.g. Google Scholar e Microsoft Acadêmico).

Um item importante nos relatórios é a lista de membros do grupo onde são mostradas informações individuais como o nível de bolsa de produtividade outorgada pelo CNPq, período de permanência do membro no grupo, e a última data de atualização do currículo Lattes.

Para fins de uma análise complementar mais apurada, todas as produções bibliográficas compiladas são armazenadas em um formato flexível denominado RIS. Um arquivo em formato RIS refere-se a um arquivo de texto ASCII onde todos os campos constituintes de um determinado artigo são indicados por duas letras, conforme descrição disponível no Reference Manager (2011). As produções bibliográficas nesse formato padrão facilitam (i) o intercâmbio de dados com diferentes bibliotecas digitais tais como IEEE Xplore, Scopus, Portal do ACM, ScienceDirect e SpringerLink, e (ii) a população de bancos de dados externos com dados relativos a produções bibliográficas de um determinado grupo.

Adicionalmente, os três tipos de grafos de colaborações computados pelo scriptLattes são armazenados em arquivos de texto ASCII, que representam as três matrizes de adjacência (MENA-CHALCO; CESAR-JR, 2009) onde, para cada linha e cada coluna,

associa-se exclusivamente um membro do grupo (membros sem alguma colaboração são definidos com valor zero na matriz de adjacência).

Salientamos que todos os tipos de relatórios são gerenciados por meio de um conjunto de parâmetros configurados na execução do programa. Isto permite manter controle exato sobre os tipos de dados a serem compilados. Por exemplo, para um determinado grupo, usualmente deseja-se: (i) a lista completa de produções bibliográficas, e (ii) o grafo de coautoria associado apenas a artigos completos publicados em periódicos.

Finalmente, destaca-se que todos os relatórios gerados são estáticos, isto é, os relatórios mantêm apenas informações obtidas durante a execução do programa (analogamente a uma fotografia da produção acadêmica no momento da execução do programa). Para um determinado grupo, as futuras atualizações nos currículos Lattes dos membros serão compiladas na frequência de re-execução do programa.

Na Figura 3 podem ser visualizadas algumas telas de exemplo de relatórios gerados automaticamente pelo scriptLattes.

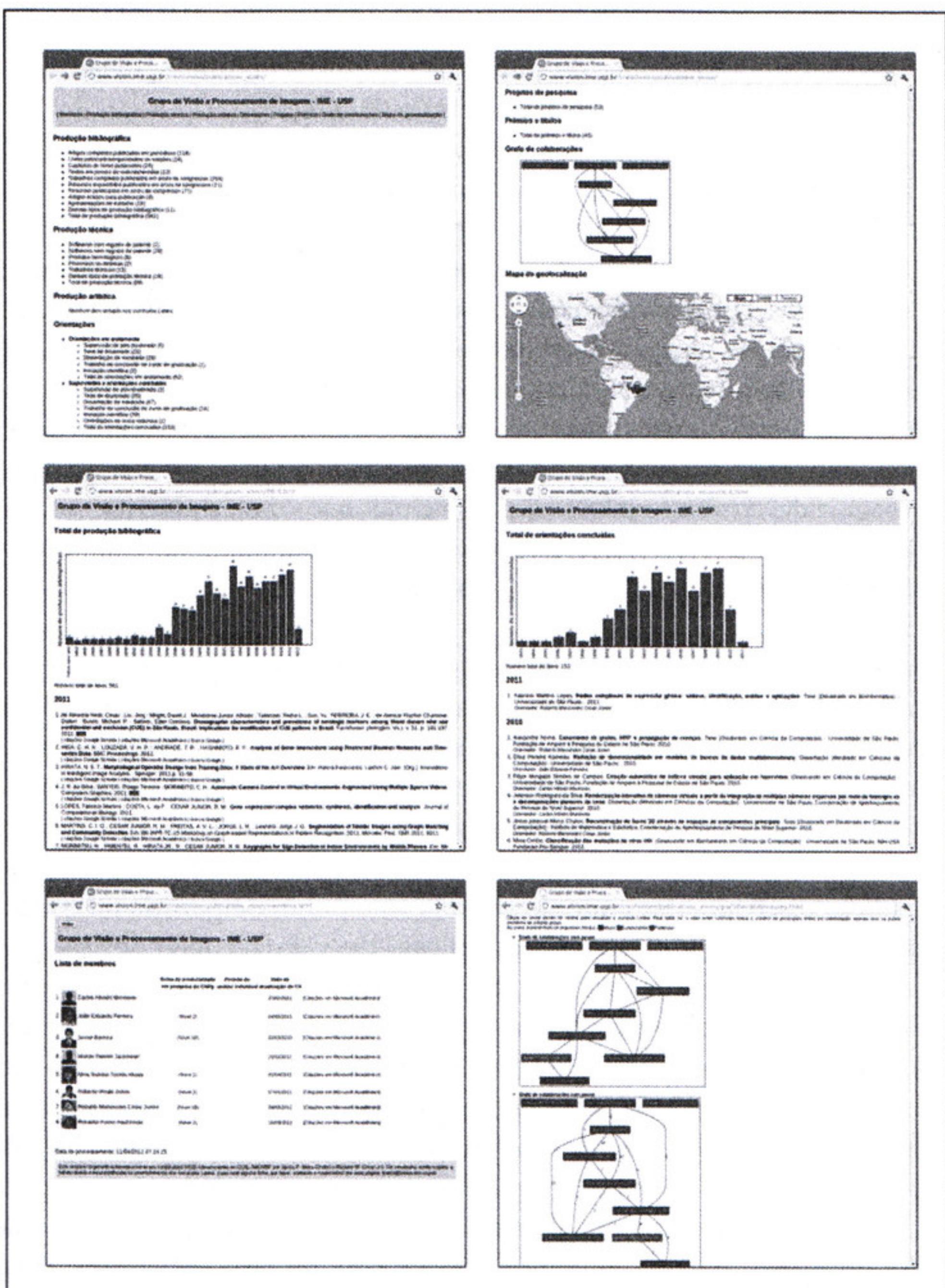


Figura 3 - Exemplo de relatórios obtidos com o scriptLattes para o grupo de Visão e Processamento de Imagens do Departamento de Ciências da Computação – USP

Os relatórios da Figura 3 estão disponíveis no site do Grupo de Visão e Processamento de Imagens do Instituto de Matemática e Estatística da Universidade de São Paulo (IME/USP).

Experiências de utilização da ferramenta

O scriptLattes foi adotado recentemente por diversas instituições de ensino e pesquisa no Brasil e a sua utilização está motivada pela necessidade de exploração automática de uma grande quantidade de currículos cadastrados na Plataforma Lattes. Um lista atualizada com as organizações e instituições de ensino e pesquisa usuárias do scriptLattes pode ser consultada no site <http://scriptlattes.sourceforge.net/links.html>. A seguir descrevemos quatro dos principais usos da ferramenta.

Criação de relatórios de produção acadêmica

Para alguns grupos de pesquisa, frequentemente é necessária a criação de relatórios de produção bibliográfica referente a alguns períodos (por ex., anual, trienal). Esses relatórios, além de detalhar as atividades acadêmicas realizadas no período, também apresentam alguns indicadores de produção em que são associadas informações como, por exemplo, o *Qualis* ou o *fator de impacto* da publicação. Nesse sentido, o scriptLattes pode ser usado como ponto de partida na criação das listas de produções acadêmicas de um grupo de pessoas cadastradas na plataforma Lattes. Embora o scriptLattes faça uma detecção de publicações iguais e/ou similares, é recomendada uma verificação cuidadosa nos relatórios automaticamente gerados, de tal forma que os possíveis erros no preenchimento dos dados nos currículos Lattes sejam corrigidos ou complementados manualmente com informações exatas.

Por outro lado, uma prática comum da ferramenta é a população de bancos de dados com as informações de produção acadêmica. Esses dados podem ser utilizados em *intranets* ou na internet para difundir o estado atual da produção acadêmica, e ajudar na tomada de decisões sobre a avaliação bibliométrica de um determinado grupo de pesquisa.

Atualmente, a principal utilização da ferramenta é a geração automática de relatórios em formato HTML para difusão da produção

acadêmica de grupos de pesquisa (como por exemplo, de programas de pós-graduação).

Criação de grafos de colaboração acadêmica

O resultado correspondente à geração automática do grafo de colaboração é um dos mais importantes dentre os relatórios gerados pelo scriptLattes. O grafo ou rede de colaboração mostra a interação de coautoria entre membros de um determinado grupo de interesse. A interação com pesquisadores não considerados no grupo em análise não é representada no grafo de colaborações, dado que o scriptLattes lida apenas com informações extraídas dos próprios currículos Lattes do grupo de interesse. Sendo assim, as colaborações com outros pesquisadores, ainda que com cadastro na plataforma Lattes, mas que não formem parte do grupo, não serão diagramadas no grafo de colaborações.

Salientamos que as matrizes de adjacência (geradas automaticamente), correspondentes aos grafos de colaboração, podem ser examinados através de ferramentas complementares de análise de redes de interação social como o Pajek (2011), Ucinet (2011) ou R_Project (2011). Desse modo, podem ser facilmente exploradas as medidas de indicadores de redes como: densidade, grau de centralidade, índice de centralização, e grau de intermediação e de proximidade (SCOTT, 2000; WASSERMAN; FAUST, 1994). Essas medidas são amplamente estudadas para: (i) caracterizar redes sociais e identificar automaticamente sub-comunidades de colaboração em grupos de pesquisa (NEWMAN; GIRVAN, 2004), (ii) estudar e caracterizar a evolução temporal das coautorias entre os membros do grupo, isto é, analisar a dinâmica de colaboração dos membros do grupo por meio de diferentes períodos de tempo (WU et al., 2009), ou (iii) correlacionar dados quantitativos de colaboração com dados qualitativos, sobre o grupo, provenientes de outras fontes de informação, a fim de examinar a tendência de atuação dos grupos sobre determinados eixos (LEYDESDORFF, 2006; 2007).

Finalmente, é importante ressaltar que também podem ser aplicadas técnicas de reconhecimento de padrões para que, através de algumas características métricas, diferentes grafos de colaboração

correspondentes a distintos grupos de pesquisa possam ser comparados a fim de ter, por exemplo, uma classificação de comportamento de perfil de publicação bibliográfica (MENA-CHALCO; CESAR-JR, 2009). Com essa formulação, grupos com produção acadêmica similar estarão próximos em um eventual espaço de características.

Criação de árvores de genealogia acadêmica

Uma extensão da utilização da ferramenta é a geração automática de árvores genealógicas individuais para cientistas/acadêmicos, cadastrados na Plataforma Lattes, através de suas relações de orientação ou supervisão concluída.

Para cada membro do grupo de interesse, pode ser gerada automaticamente a *ascendência* (pais) e *descendência* (filhos) de orientação acadêmica. Caso o identificador Lattes do orientador/coorientador ou do aluno seja identificado no currículo Lattes, o nó é expandido por mais um nível (a quantidade de níveis pode ser limitada por um valor informado pelo usuário).

Essa estratégia de elaboração de árvores de genealogia acadêmica pode ser explorada recursivamente para manter um banco de dados com as relações de orientação acadêmica Lattes, similar ao do projeto de genealogia matemática da Sociedade Americana de Matemática (AMS, 2011). Acreditamos que trabalhos nesta linha têm um grande potencial para análises automáticas de inter-relações de orientação associadas para todas as áreas de pesquisa no Brasil.

De maneira similar à criação automática de árvores de genealogia acadêmica Lattes, redes de colaboração podem ser criadas entre todos os coautores de um determinado grupo. Assim, o número de nós no grafo criado não seria limitado pela quantidade de membros e sim pela quantidade real de colaboradores cadastrados na Plataforma Lattes. Essa abordagem, embora requeira diversas consultas de currículos à plataforma Lattes, apresentaria um panorama macro da influência de colaboração entre pesquisadores. Como resultado dessa compilação de inter-relações, pode ser definido um valor numérico que represente a “Distância Lattes” entre dois pesquisadores, ou seja, o número mínimo de arestas que tem que ser percorridas para ligar esses 2 pesquisadores, análogo ao Número Erdos que representa a distância

de coautoria entre um qualquer pesquisador e Paul Erdos (BATAKELJ; MRVAR, 2000).

Análise da distribuição geográfica de pesquisadores

O impacto da formação acadêmica de um determinado grupo, através da localização espacial ou geográfica, também é outra prática comum da utilização do scriptLattes. As relações de orientação podem ser examinadas geograficamente, tendo assim uma noção de distribuição espacial dos membros do grupo e dos alunos formados.

A distribuição geográfica de um grupo de interesse pode ser analisada para obter informações como, por exemplo, o estado ou região que atrai mais os alunos formados, isto é, a influência de um estado ou região sobre outras, por exemplo. Essas informações potencialmente podem produzir estatísticas relevantes, desde que se mantenha uma adequada normalização nos dados de geolocalização.

Certamente, existe um grande desafio em compreender como a distribuição geográfica é conceituada, medida e normalizada (PITBLADO; PONG, 1999). Acreditamos que o tratamento dos dados, como o realizado pelo scriptLattes, é uma abordagem plausível a ser considerada para investigar indicadores demográficos tanto de orientação quanto de formação acadêmica no Brasil.

Por fim, a ferramenta pode ser modificada para manter uma representação (i) do grafo de colaboração e (ii) das árvores de genealogia acadêmica, conjuntamente com as localizações geográficas dos pesquisadores, de maneira que as informações tanto de produção acadêmica, quanto de geolocalização sejam fusionadas, permitindo assim uma possível descoberta de informação de produção de membros do grupo através de suas inter-relações (ENKHSAIKHAN; LIU; REYNOLDS, 2008).

Aspectos de implementação computacional

Inicialmente, o scriptLattes foi desenvolvido em 2005 na linguagem de programação Perl. Entretanto, a versão de 2011 foi reprogramada inteiramente na linguagem Python. O código fonte de ambas as versões é distribuído na modalidade de software livre sob a

licença GNU-GPL que, entre outras liberdades, permite executar o programa para qualquer propósito, estudar seu funcionamento e realizar possíveis adaptações para determinadas necessidades. A nova versão do scriptLattes em Python permite uma rápida adaptação/modificação dos procedimentos para diversas finalidades (sugeridas principalmente para atividades acadêmicas ou de pesquisa), pois foi criada com estruturas de dados simples, módulos padrão da linguagem de programação, e seguindo o paradigma da Programação Orientada a Objetos. É importante frisar que a distribuição do código fonte permite também a execução do programa sob diferentes sistemas operacionais (e.g. Windows, Linux e MacOS) desde que os módulos requeridos sejam corretamente instalados no próprio sistema operacional.

Finalmente, o tempo de execução do programa dependerá, além do tempo de conexão com as plataformas Lattes e *Google Maps*, do número total de produções acadêmicas produzidas e cadastradas pelo grupo, e não do número de membros considerados no grupo. Nesse sentido, compilar os dados acadêmicos de um grupo pequeno com muitas produções bibliográficas será mais demorado do que compilar os dados acadêmicos de um grupo médio ou grande com pouquíssimas produções bibliográficas.

Considerações finais

O scriptLattes é um programa, ainda em desenvolvimento, que auxilia principalmente na compilação ou coleta de dados de currículos Lattes que, tipicamente, é difícil de obter de forma manual para grupos de médio ou grande porte. O objetivo deste texto foi descrever as principais características da ferramenta, assim como apresentar algumas experiências de sua utilização sobre um conjunto de dados extremamente valioso, entretanto pouco explorado, como é a Plataforma Lattes.

Salientamos que a utilização de ferramentas automáticas similares à apresentada, tais como DBLP, CiteSeer, Google Scholar, Microsoft Academic Search, e ArnetMiner (TANG et al., 2008) são cada vez mais necessárias, pois existe um volume crescente de dados de produção acadêmica e científica que devem ser corretamente

computados e explorados visualmente de forma efetiva por meio de métodos computacionais (KEIM, 2002).

Embora seja louvável a tarefa de compilação automática de dados acadêmicos para grupos de pesquisa, deve-se perceber que os resultados apenas refletem os dados cadastrados na Plataforma Lattes. Consequentemente, dados cadastrados de forma incorreta e/ou incompleta nos currículos Lattes também permanecerão incorretos e/ou incompletos nos relatórios de compilação gerados pelo scriptLattes.

Como trabalhos futuros, pretende-se explorar outros dados disponíveis nos currículos Lattes como, por exemplo, a formação acadêmica/titulação e a identificação das áreas de atuação. Essas informações são importantes para a avaliação acadêmica de grupos, possibilitando identificar uma correspondência com o grau de internacionalização do grupo (por exemplo, quais foram os países em que os membros do grupo se titularam). Adicionalmente, pretende-se implementar uma estratégia de atualização incremental de todos os relatórios gerados para currículos Lattes correspondentes a novos membros do grupo. Com isto, a nova informação será acrescentada aos relatórios, sem ter a necessidade de processar os currículos Lattes de todos os membros do grupo. Finalmente, pretende-se investir tanto na adoção de sistemas de visualização eficiente de grafos complexos que lidam com quantidades grandes de nós, quanto na utilização de novas medidas para representação de grafos de colaborações como, por exemplo, a influência na colaboração (TANG; YANG, 2009).

Finalmente, é válido ressaltar que o scriptLattes não está vinculado ao CNPq e por decorrência esta agência de fomento à pesquisa científica e tecnológica não é responsável por nenhuma assessoria técnica sobre esta ferramenta. A ferramenta é o resultado de um esforço independente realizado com o único intuito de auxiliar as tarefas mecânicas de compilação ou coleta de informações publicamente cadastradas nos Currículos Lattes e as dúvidas técnicas podem ser encaminhadas ao primeiro autor desse texto.

Referências

- AMERICAN Mathematical Society. Mathematical genealogy project. Disponível em: <<http://genealogy.math.ndsu.nodak.edu/index.php>>. Acesso em jan. 2011.
- AMORIN, C. V. Curriculum vitae organization: the Lattes software platform. **Pesquisa Odontológica Brasileira**, v. 17, n. 1, p. 18–22, 2003.
- BARABASI, A. L.; ALBERT, R. Emergence of scaling in random networks. **Science**, v. 286, n. 5439, p. 509–512, 1999.
- BATAGELJ, V.; MRVAR, A. Some analyses of Erdos collaboration graph. **Social Networks**, v. 22, n. 2, p. 173–186, 2000.
- ENKHSAIKHAN, M.; LIU, W.; REYNOLDS, M. Geographical and temporal visualisation of social relationships. In: PACIFIC ASIA CONFERENCE ON INFORMATION SYSTEMS, 12., 2008, Suzhou. **Proceedings...** 2008. Disponível em: <http://www.pacis-net.org/file/2008/PACIS2008_Camera-Ready_Paper_243.pdf>. Acesso em: jan. 2011.
- GRUPO de Visão e Processamento de Imagens. IME-USP. Disponível em: <http://www.vision.ime.usp.br/creativision/publications_vision>. Acesso em: jan. 2011.
- KANG, I. S. et al. On co-authorship for author disambiguation. **Information Processing and Management**, v. 45, n. 1, p. 84–97, 2009.
- KEIM, D. A. Information visualization and visual data mining. **IEEE Transactions on Visualization and Computer Graphics**, v. 7, n. 1, p. 100-107, 2002.
- KLINK, S. et al. Analysing social networks within bibliographical data. 7th International Conference on Database and Expert Systems Applications. **Lecture Notes in Computer Science**, v. 4080, p. 234–243, 2006.
- KOUZES, R. T. et al. The changing paradigm of dataintensive computing. **Computer**, v. 42, n. 1, p. 26–34, 2009.
- LEYDESDORFF, L. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. **Journal of the American Society for Information Science and Technology**, v. 58, p. 1303–1319, 2007.
- LEYDESDORFF, L. Can scientific journals be classified in terms of aggregated journal-journal citation relations using the journal citation reports?. **Journal of the American Society for Information Science & Technology**, v. 57, n. 5, p. 601-613, 2006.
- LIU, X. et al. Co-authorship networks in the digital library research community. **Information Processing and Management**, v. 41, n. 6, p. 1462-1480, 2005.
- MAIA, M. F.; CAREGNATO, S. E. Co-autoria como indicador de redes de colaboração científica. **Perspectivas em Ciência da Informação**, v. 13, n. 2, p. 18–31, 2008.

MENA-CHALCO, J. P.; CESAR-JR, R. M. scriptLattes: an open-source knowledge extraction system from the lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31–39, 2009.

NAVARRO, G. A guided tour to approximate string matching. **ACM Computing Surveys**, v. 33, n. 1, p. 31–88, 2001.

NEWMAN, M. E. J.; GIRVAN, M. Finding and evaluating community structure in networks, **Physical Review E**, v. 69, n. 2, p. 026113(15), 2004.

NICHOLSON, S. The basis for bibliomining: frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. **Information Processing and Management**, v. 42, n. 3, p. 785–804, 2006.

PAJECK. **Program for large network analysis**. Disponível em: <<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>>. Acesso em: jan. 2011.

PENG, F.; MCCALLUM, A. Information extraction from research papers using conditional random fields. **Information Processing and Management**, v. 42, n. 4, p. 963–979, 2006.

PITBLADO, J. R.; PONG, R. W. **Geographic distribution of physicians in Canada**. Sudbury, Laurentian University Centre for Rural and Northern Health Research, 1999.

REFERENCE Manager. RIS format specifications. Disponível em: <http://www.refman.com/support/risformat_intro.asp>. Acesso em: jan. 2011.

SCOTT, J. **Social network analysis**: a handbook. 2.ed. London: Sage, 2000.

SCRIPTLATTES. Disponível em: <http://scriptlattes.sourceforge.net> Acesso em jan. 2011.

TANG, J. et al. Arnetminer: extraction and mining of academic social networks. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 14., 2008. **Proceedings** ... p. 990–998.

TANG, J.; YANG, Z. Social influence analysis in large-scale social networks. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 15., 2009, **Proceedings**... p. 807–816.

THE R PROJECT for Statistical Computing. Disponível em: <<http://www.r-project.org/>>. Acesso em: jan. 2011.

TOMITA, M. **Current issues in parsing technology**. Boston: Kluwer Academic Publishers, 1991.

UCINET Software. Disponível em: <<http://www.analytictech.com/ucinet>>. Acesso em: jan. 2011.

WASSERMAN, S.; FAUST, K. **Social network analysis**. Cambridge: Cambridge University Press, 1994.

WU, B. et al. Characterizing the evolution of collaboration network. In: ACM WORKSHOP ON SOCIAL WEB SEARCH AND MINING, 2., 2009. *Proceedings* ... p. 33–40.

YE, N. **The handbook of data mining**. Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers, 2003.