LUCIANO ANTONIO DIGIAMPIETRI

Análise da Rede Social Acadêmica Brasileira

São Paulo

2015

LUCIANO ANTONIO DIGIAMPIETRI

Análise da Rede Social Acadêmica Brasileira

Versão corrigida.

Texto sistematizando parte da obra do candidato apresentado à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Livre-Docente.

Área de Concentração: Informação e tecnologia

São Paulo
2015

Dedico este trabalho a minha esposa e a meu filho.

Agradecimentos

Agradeço

A Deus por me dar todas as oportunidades, amigos e desafios que me permitiram desenvolver este trabalho.

A minha família por todo o apoio e paciência.

A meus ex-orientadores que guiaram meus primeiros passos na pesquisa acadêmica e estão sempre me inspirando.

Aos amigos e colegas de trabalho que tanto se dedicam à construção e à consolidação desta nova unidade da USP e que muito colaboram com todos os projetos de pesquisa em que participo.

Aos meus orientandos que dividem diariamente comigo as atividades de pesquisa.

A todos os meus alunos que são a grande motivação para meu trabalho acadêmico.

A Universidade de São Paulo por fornecer infraestrutura e fomento para o meu desenvolvimento acadêmico.

A FAPESP, CNPq, CAPES e MEC pelo apoio financeiro.

"Se enxerguei mais longe, foi por estar sobre os ombros de gigantes."

(Isaac Newton)

RESUMO

DIGIAMPIETRI, Luciano Antonio. **Análise da Rede Social Acadêmica Brasileira**. 2015. 160 f. Tese (Livre Docência) - Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2015.

No Brasil existe uma gama muito grande de informações disponíveis sobre a produção bibliográfica e outras atividades acadêmicas. Apesar deste grande conjunto de dados, existem diversos desafios relacionados à efetiva extração de conhecimento a partir dele de forma a possibilitar, por exemplo, a criação de políticas científicas nacionais eficientes e adequadas à diversidade brasileira. Este trabalho apresenta diferentes iniciativas de pesquisa realizadas pelo autor em colaboração com orientados e colegas de trabalho tanto para a caracterização de parte da rede social acadêmica brasileira, quanto do uso das informações acadêmicas disponíveis para o teste e a validação de novas estratégias para tratar aspectos específicos da análise de redes sociais, como predição de relacionamentos ou análise de tendências.

Palavras-chave: Análise de Redes Sociais; Bibliometrica; Redes Sociais Acadêmicas

ABSTRACT

DIGIAMPIETRI, Luciano Antonio. **Analysis of the Brazilian academic social network**. 2015. 160 p. Thesis - School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2015.

There is in Brazil a wide range of information available on the bibliographic production and other academic activities. Despite this large data set, there are several challenges related to the effective extraction of knowledge from it in order to enable, for example, the creation of effective national science policy, adequate for the Brazilian diversity. This work presents different research initiatives carried out by the author in collaboration with advisees and co-workers for both the characterization of the Brazilian academic social network, and the use of academic information available for testing and validating new strategies to address specific aspects of the social network analysis such as link prediction and trend analysis.

Keywords: Social Network Analysis; Bibliometrics; Academic Social Networks

Lista de figuras

Lista de tabelas

Sumário

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

As atividades acadêmicas, assim como grande parte das atividades humanas, são construídas ao redor de relacionamentos. Assim, a atuação de um indivíduo não pode ser plenamente entendida sem a devida contextualização deste em sua rede de relacionamentos. O conjunto formado pelos indivíduos e suas ligações é denominado rede social. A área dedicada ao estudo, caracterização e visualização dos dados relacionados a redes sociais é chamada de análise de redes sociais. Nesta área, considera-se que as relações entre os indivíduos são tão importantes ou mesmo mais importantes para caracterizar um conglomerado social do que as características específicas de cada indivíduo (WASSERMAN; FAUST, 2009).

Atualmente, uma grande quantidade de informações referentes à produção científica está disponível na Web: publicações científicas, informações sobre projetos de pesquisa e mesmo currículos de pesquisadores.

No que tange aos dados referentes à pesquisa, o Brasil apresenta uma característica peculiar: a existência de um cadastro nacional de currículos de pesquisadores, a Plataforma Lattes, que congrega informações sobre publicações, orientações, projetos de pesquisa, entre outras. Esse grande volume de informações ainda não foi amplamente usado e, tipicamente, é consultado para avaliar (ou verificar dados) de pesquisadores individualmente ou de alguns grupos de pesquisadores (por exemplo, docentes credenciados em programas de pós-graduação). Adicionalmente, as informações dos currículos podem ser enriquecidas com informações de outras bases bibliográficas ou bases que contenham contadores de citações para a análise da inserção da produção nacional em bases internacionais.

Estudos clássicos sobre currículos costumam abordar questões referentes à produção bibliográfica de pesquisadores (PRITCHARD, 1969; TAGUE-SUTCLIFFE, 1992; CALLON et al., 1995; SPINAK, 1998) ou questões de trajetória na carreira, mobilidade e mapeamento da capacidade coletiva (CANIBANO; BOZEMAN, 2009). Porém, os currículos de pesquisadores podem ser cruzados de forma a estabelecer relações entre os mesmos. O conjunto formado por pesquisadores e suas relações pode ser visto como uma Rede Social Acadêmica (*Academic Social Network*), a qual pode ser estudada e caracterizada utilizando teoria dos grafos (BERKOWITZ, 1982; WASSERMAN; GALASKIEWICZ, 1994; WASSERMAN; FAUST, 2009; SCOTT, 2009; BREIGER, 2004; ULRIK; ERLEBACH, 2005; LEMIEUX; OUIMET, 2008; POBLACION; MUGNAINI; RAMOS, 2009; PRELL, 2012).

A pesquisa apresentada neste documento visa a combinar características da análise bibliométrica com a análise de redes sociais com diferentes objetivos. Por um lado, ambas as análises são utilizadas para a caracterização da rede social acadêmica, apresentando uma visão geral da rede formada pelos possuidores dos currículos Lattes e uma visão mais detalhada de alguns grupos específicos Por outro lado, a grande quantidade e riqueza da informação disponibilizada pela Plataforma Lattes pode ser utilizada para o teste e a validação de novas estratégias (novos algoritmos ou combinação de algoritmos existentes) para resolver diferentes problemas da análise de redes sociais.

Os resultados apresentados neste trabalho estão contextualizados dentro do Grupo de Análise de Redes Sociais e Cientometria [1]. Este grupo interdisciplinar de pesquisadores da Escola de Artes, Ciências e Humanidades da USP iniciou alguns estudos em colaboração, com vistas à geração de indicadores dos mais diversos aspectos do currículo acadêmico dos pesquisadores brasileiros, tendo a Plataforma Lattes como fonte inicial de informação. Atualmente, o grupo vem expandindo seu olhar para outras fontes, tais como: bases de dados de publicações científicas (como Web of Science[2], SciELO[3] e DBLP[4]) e bases de dados institucionais (como Dedalus[5] e Tycho[6]).

Entre os desafios previstos para este projeto estão: desenvolvimento de algoritmos para caracterização e identificação de tendências na produção científica nacional; predição/sugestão de possíveis coautorias; identificação de especialistas nas diferentes áreas e subáreas do conhecimento; avaliação da importância da geolocalização dos pesquisadores em relação às suas colaborações científicas e recomendação de artigos científicos.

1.1 Objetivos

Este documento tem por objetivo geral congregar e descrever a pesquisa desenvolvida pelo docente ao longo dos últimos cinco anos sobre o tema: Análise da Rede Social Acadêmica Brasileira.

A fim de atingir o objetivo geral, os seguintes objetivos específicos foram definidos:

1 http://dgp.cnpq.br/dgp/espelhogrupo/9125239221851493
2 http://wokinfo.com/
3 http://www.scielo.org/
4 dblp.uni-trier.de
5 http://dedalus.usp.br/
6 https://uspdigital.usp.br/tycho/index.jsp

1. Descrição dos conceitos básicos da área contextualizando a pesquisa desenvolvida;

2. Detalhamento dos principais resultados da pesquisa.

1.2 Metodologia

A metodologia para o desenvolvimento da pesquisa apresentada neste trabalho foi composta de diversos ciclos, contento as seguintes atividades: estudo teórico do assunto abordado, baseado principalmente em artigos científicos e livros, sobre técnicas de análise de redes sociais, predição de relacionamentos, mineração de textos, identificação de tendências e análise bibliométrica/cientométrica de grupos; com base no estudo foram especificadas, desenvolvidas, testadas e validadas ferramentas para a realização de cada uma das atividades deste projeto, comparando-se os resultados obtidos com os disponíveis na literatura; as ferramentas desenvolvidas foram, então, aplicadas em estudos de caso reais e os resultados dessa aplicação foram analisados e publicados.

Ao longo dos capítulos deste documento serão detalhados alguns aspectos metodológicos referentes ao desenvolvimento de cada parte da pesquisa realizada.

1.3 Publicações

Esta seção apresenta a lista de produções bibliográficas do autor dentro do contexto da análise da rede social acadêmica brasileira, incluindo os trabalhos relacionados especificamente a análises bibliométricas. Os trabalhos apresentados variam de artigos completos aceitos para publicação até resumos/pôsteres apresentados em eventos.

As tabelas 1, 2 e 3 apresentam as publicações da seguinte forma: a primeira coluna contém a referência à publicação; a segunda coluna contém o título, por fim, a última coluna apresenta o principal assunto tratado pelo respectivo trabalho.

1.4 Orientações

Esta seção apresenta as orientações realizadas nos últimos cinco anos (concluídas ou em andamento), cujos projetos estão relacionados ao tema de pesquisa apresentado neste documento.

Tabela 1 – Lista das publicações - periódicos

Artigos aceitos para publicação em periódicos		
Referência	**Título**	**Principal Assunto**
(CHAGAS; PEREZ-ALCAZAR; DIGIAMPIETRI, 2015)	Algoritmo de classificação de especialistas em áreas na base de currículos Lattes	identificação de especialistas
(DIGIAMPIETRI et al., 2015b)	Análise da evolução das relações de coautoria nos programas de pós-graduação em computação no Brasil	dinâmica da rede social acadêmica
(DIGIAMPIETRI et al., 2015a)	Extração, caracterização e análises de dados de currículos Lattes	obtenção e organização dos dados
(TUESTA et al., 2015a)	Análise comparativa da produtividade dos pares orientador-orientado em Ciência da Computação	relação orientador-orientado
Artigos publicados em periódicos		
Referência	**Título**	**Principal Assunto**
(TUESTA et al., 2015b)	Analysis of an Advisor-Advisee Relationship: An Exploratory Study of the Area of Exact and Earth Sciences in Brazil	relação orientador-orientado
(DIGIAMPIETRI et al., 2014)	BraX-Ray: An X-Ray of the Brazilian Computer Science Graduate Programs	análise de grupos
(DIGIAMPIETRI et al., 2014b)	Análise macro das últimas atualizações dos Currículos Lattes	análise da atualização dos dados
(MENA-CHALCO et al., 2014)	Brazilian bibliometric coauthorship networks	análise de grupos
(MUGNAINI; DIGIAMPIETRI; MENA-CHALCO, 2014a)	Comunicação científica no Brasil (1998-2012): indexação, crescimento, fluxo e dispersão	análise bibliométrica
(TRUCOLO; DIGIAMPIETRI, 2014a)	Análise de Tendências da Produção Científica Nacional da Área de Ciência da Computação	análise de tendências
(BRITO; DIGIAMPIETRI, 2013)	Análise de Tendências da Produção Científica Nacional da Área de Ciência da Computação	recomendação de conteúdo
(MENA-CHALCO; DIGIAMPIETRI; OLIVEIRA, 2012)	Perfil de produção academica dos programas brasileiros de pós-graduação em Ciência da Computação nos trienios 2004-2006 e 2007-2009	análise de grupos
(MUGNAINI et al., 2012)	Normalização de nomes de autores em fontes de informação institucionais: proposta de um método automático de verificação de erros	obtenção e organização dos dados
(DIGIAMPIETRI; SILVA, 2011)	A Framework for Social Network of Researchers Analysis	análise de grupos

Fonte: Digiampietri (2015)

A tabela 4 contém a lista dos alunos orientados, título de seus projetos e período de desenvolvido.

Tabela 2 – Lista das publicações - artigos completos em anais

Artigos completos publicados em anais de eventos		
Referência	**Título**	**Principal Assunto**
(DIGIAMPIETRI et al., 2015)	Um Sistema de Predição de Relacionamentos em Redes Sociais	predição de relacionamentos
(MUGNAINI; DIGIAMPIETRI, 2015)	The Brazilian national impact: movement of journals between Bradford Zones of production and consumption	análise bibliométrica
(MUGNAINI; DIGIAMPIETRI; MENA-CHALCO, 2014b)	Comunicação científica no Brasil (1998-2012): infraestrutura nacional e internacionalização	análise bibliométrica
(DIGIAMPIETRI et al., 2014a)	Análise da Atualização dos Currículos Lattes	análise da atualização dos dados
(TRUCOLO; DIGIAMPIETRI, 2014b)	Uma Revisão Sistemática acerca das Técnicas de Identificação e Análise de Tendências	análise de tendências
(DIGIAMPIETRI et al., 2014a)	Análise da Rede de Relacionamentos dos Doutores Brasileiros	análise de grupos
(DIGIAMPIETRI et al., 2014b)	Análise da Rede dos Doutores que Atuam em Computação no Brasil	análise de grupos
(DIGIAMPIETRI; PERES; SILVA, 2014)	Rede de Relacionamentos Brasileira de Inteligência Artificial e Computacional	análise de grupos
(DIGIAMPIETRI; MENA-CHALCO, 2013)	Correlation among the scientific production, supervisions and participation in defense examination committees in the Brazilian physicists community	análise de grupos / bibliométrica
(MELO-MINARDI et al., 2013)	Caracterização dos programas de pós-graduação em Bioinformática no Brasil	análise de grupos
(DIGIAMPIETRI; MUGNAINI; ALVES, 2013)	Análise da participação dos orientandos na produção dos orientadores: um estudo de caso em Ciência da Computação	relação orientador-orientado
(DIGIAMPIETRI; SANTIAGO; ALVES, 2013)	Predição de coautorias em redes sociais acadêmicas: um estudo exploratório em Ciência da Computação	predição de relacionamentos
(MIYATA; KANO; DIGIAMPIETRI, 2013a)	Combinando mineração de textos e análise de redes sociais para a identificação das áreas de atuação de pesquisadores	identificação de áreas de pesquisa
(DIGIAMPIETRI et al., 2012a)	Minerando e Caracterizando Dados de Currículos Lattes	obtenção e organização dos dados
(MENA-CHALCO; DIGIAMPIETRI; CESAR-JUNIOR, 2012)	Caracterizando as redes de coautoria de currículos Lattes	análise de grupos
(DIGIAMPIETRI et al., 2012b)	Dinâmica das Relações de Coautoria nos Programas de Pós-Graduação em Computação no Brasil	dinâmica da rede
(TUESTA et al., 2012)	Análise temporal da relação orientador-orientado: um estudo de caso sobre a produtividade dos pesquisadores doutores da área de Ciência da Computação	relação orientador-orientado
(MENA-CHALCO; DIGIAMPIETRI; OLIVEIRA, 2012)	Perfil de produção acadêmica dos programas brasileiros de pós-graduação em Ciencia da Computação nos triênios 2004-2006 e 2007-2009	análise de grupos
(MUGNAINI; LEITE; LETA, 2011)	Fontes de Informação para Análise de internacionalização da Produção Científica Brasileira	análise bibliométrica
(PEREZ-ALCAZAR et al., 2011)	Avaliação de Redes de Inovação usando uma ferramenta baseada em redes sociais - caso Brasileiro de Nanotecnologia	análise de grupos

Fonte: Digiampietri (2015)

1.5 Organização do documento

O restante deste documento está organizado da seguinte forma. O capítulo 2 contém a descrição dos conceitos básicos utilizados ao longo do trabalho, contextualizando os resultados obtidos e sumarizando os trabalhos correlatos. O capítulo 3 apresenta as ferramentas desenvolvidas para a obtenção e organização dos dados. No capítulo 4 são detalhadas as abordagens utilizadas para a resolução de entidades. O capítulo 5 apresenta

Tabela 3 – Lista das publicações - outras

Capítulo de livro		
Referência	**Título**	**Principal Assunto**
(LIMA; DIGIAMPIETRI, 2014)	Enriquecendo base de dados de currículos Lattes	obtenção e organização dos dados
Resumos expandidos publicados em anais		
Referência	**Título**	**Principal Assunto**
(DIGIAMPIETRI; BARBOSA; LINDEN, 2015)	Desambiguação de nomes em redes sociais acadêmicas: Um estudo de caso usando DBLP	obtenção e organização dos dados
(DIGIAMPIETRI; MARUYAMA, 2014)	Predição de Novas Coautorias na Rede Social Acadêmica dos Programas Brasileiros de Pós-Graduação em Ciência da Computação	predição de relacionamentos
Resumos publicados em anais		
Referência	**Título**	**Principal Assunto**
(MIYATA; KANO; DIGIAMPIETRI, 2013b)	Uso de Mineração de Textos para a Identificação das Áreas de Atuação de Pesquisadores	identificação de áreas de pesquisa
(KANO; MIYATA; DIGIAMPIETRI, 2013)	Uso de Mineração de Textos para Análise de Características da Produção Científica Nacional	identificação de áreas de pesquisa
(ALVES; DIGIAMPIETRI, 2013)	Análise de Redes Sociais: Desenvolvimento de Ferramentas para a Análise da Comunidade Científica Brasileira	obtenção e organização dos dados
(LIMA; DIGIAMPIETRI, 2013)	Enriquecendo bases de dados de currículos Lattes	obtenção e organização dos dados
(SILVA; DIGIAMPIETRI, 2012)	Análise de Redes Sociais de Pesquisadores Baseada em Dados da Plataforma Lattes	análise de grupos
(FREIRE; DIGIAMPIETRI, 2011)	Desenvolvimento de Sistema de Recomendação de Artigos Científicos	recomendação de conteúdo

Fonte: Digiampietri (2015)

as duas principais estratégias utilizadas nesta pesquisa para a análise de grupos e alguns dos resultados obtidos. No capítulo 6 são apresentados os resultados referentes à predição de relacionamentos em redes sociais. O capítulo 7 sumariza alguns resultados obtidos na identificação automática de áreas de atuação, análise de tendências e análise da relação orientador-orientado. Por fim, o capítulo 8 contém as conclusões e algumas direções para trabalhos futuros.

Tabela 4 – Orientações

Mestrado		
Aluno	**Título**	**Período**
Jamison José da Silva Lima	Identificação de especialistas em redes sociais acadêmicas	2015 –
Lênin Ferreira Barbosa	Avaliação de grupos de pesquisa baseado em análise de redes sociais e cientometria	2015 –
Caio Cesar Trucolo	Análise de tendências em redes sociais acadêmicas	2013 –
William Takahiro Maruyama	Predição de coautorias em redes sociais acadêmicas	2013 –
Arthur Patricio Grava	Sistema de recomendação de artigos para pesquisadores da plataforma Lattes	2013 –
Iniciação Científica		
Caio Margutti Alves	Análise de Redes Sociais: Desenvolvimento de Ferramentas para a Análise da Comunidade Científica Brasileira	2013–2014
Rodrigo Antonio de Freitas Vieira	Análise de Redes Sociais de Pesquisadores Cadastrados na Plataforma Lattes	2013–2014
Fernando Mendes Stefanini	Desenvolvimento de sistema de recomendação de atividades em workflows	2013–2014
Jamison José da Silva Lima	Desenvolvimento de Ferramentas para a Análise de Redes Sociais Acadêmicas	2013–2014
Jamison José da Silva Lima	Enriquecendo Bases de Dados de Currículos Lattes	2012–2013
Bruno Kazuhiro Oliveira Miyata	Uso de mineração de textos para a identificação das áreas de atuação de pesquisadores	2012–2013
Vitor Yudi Kano	Uso de mineração de textos para análise de características da produção científica nacional	2012–2013
Gabriela Scardine Silva	Análise de Redes Sociais de Pesquisadores Baseada em Dados da Plataforma Lattes	2011–2012

Fonte: Digiampietri (2015)

2 Conceitos básicos e trabalhos correlatos

Rede social é uma estrutura composta por indivíduos (pessoas ou organizações) que são conectadas por um ou mais tipos de relações, por exemplo, amizade, crença ou trabalho. O conceito de redes sociais surgiu há séculos nas áreas de antropologia social e sociologia, porém, apenas nas últimas décadas, os estudos em redes sociais se intensificaram analisando tanto as características individuais de seus componentes, como também as características estruturais da rede (LEMIEUX; OUIMET, 2008).

A análise de redes sociais vem sendo utilizada em diferentes tipos de aplicação (WASSERMAN; GALASKIEWICZ, 1994). Até o início da década de 1980, a maioria da pesquisa era realizada analisando-se pequenos grupos cujos dados tipicamente eram obtidos por entrevistas ou questionários (WASSERMAN; FAUST, 2009). A partir dessa década, diferentes registros, incluindo aqueles no formato digital, começaram a fazer parte da análise de redes sociais formadas por milhares ou milhões de indivíduos e, cada vez mais, com o apoio computacional.

Há muitos estudos desenvolvidos na área de análise de redes sociais. Dois tipos estão mais relacionados ao presente projeto: os trabalhos relacionados à caracterização das redes por meio do cálculo de métricas utilizando, por exemplo, teoria dos grafos/teoria de redes (BERKOWITZ, 1982; BREIGER, 2004; ULRIK; ERLEBACH, 2005) e os trabalhos que analisam redes sociais formadas a partir de dados da Plataforma Lattes.

Um dos métodos mais comuns para a representação computacional de redes sociais é utilizando grafos. Neles, cada indivíduo da rede social é representado como um nó (ou vértice), e cada relação entre indivíduos é representada como uma aresta. A decisão se o grafo será direcionado ou não, depende do tipo de relação que se pretende representar.

Há diversos conceitos e métricas relacionados a grafos que são úteis para a análise de redes sociais, uma breve descrição destes conceitos é apresentada ao longo deste capítulo. A escolha das métricas que serão aplicadas ao estudo de cada rede social está diretamente relacionada com objetivos específicos do estudo a ser realizado (SCOTT, 2009).

Dados oriundos da Plataforma Lattes têm sido utilizados como a principal fonte de dados de diversas pesquisas acadêmicas, em particular na última década. Estas pesquisam variam da tentativa de oferecer uma visão geral sobre toda a produção científica brasileira (LEITE; MUGNAINI; LETA, 2011) ou de grupos de pesquisa específicos (ARRUDA et al., 2009; WAINER; VIEIRA, 2013; COSTA; PEDRO; MACEDO, 2013), ou mesmo pesquisas

que propõem ferramentas para auxiliar na extração, organização e visualização de dados da Plataforma Lattes (ALVES; YANASSE; SOMA, 2011a; ALVES; YANASSE; SOMA, 2011b; MENA-CHALCO; CESAR-JUNIOR, 2009).

Este capítulo apresenta os principais conceitos utilizados ao longo do texto, bem como uma contextualização dos trabalhos correlatos.

2.1 Métricas da análise de redes sociais

Nesta seção, são apresentadas algumas das métricas utilizadas na análise de redes sociais. As redes sociais são tipicamente representadas como grafos, assim serão revisados de maneira bastante sucinta conceitos básicos relacionados a grafos. Adicionalmente, uma contextualização sobre o uso desses conceitos na análise de redes sociais e as interpretações tipicamente dadas a estes nesse tipo de análise será apresentada.

A bibliografia principal utilizada no presente trabalho sobre teoria dos grafos foram os livros de Szwarcfiter (1986), Cormen et al. (2001), Newman (2003b) e Diestel (2006). A bibliografia para a análise das redes sociais, em termos de quais medidas utilizar e como interpretá-las, foi composta principalmente pelos livros de Wasserman e Faust (2009), Wasserman e Galaskiewicz (1994), Scott (2009), Lemieux e Ouimet (2008), Poblacion, Mugnaini e Ramos (2009) e Prell (2012).

Um grafo $G = (V, E)$ é composto por um conjunto de vértices ou nós V e um conjunto de arestas E, sendo que cada aresta representa uma conexão entre dois elementos do conjunto V. Há duas maneiras padrão de representar um grafo: um conjunto de listas de adjacências ou uma matriz de adjacências (CORMEN et al., 2001). A primeira tem a vantagem de ser mais compacta, especialmente útil para grafos *esparsos*. Já a segunda pode ser preferível quando o grafo for *denso* e/ou quando se precisa realizar um processamento no grafo que seja facilitado pela existência desta matriz (por exemplo, identificar rapidamente se uma aresta está conectando dois vértices quaisquer) (CORMEN et al., 2001).

Um *grafo esparso* é aquele cujo número de arestas (denotado como $|E|$) é muito menor do que o número de vértices ao quadrado ($|V|^2$). Por outro lado, um *grafo denso* é aquele cujo número de arestas é próximo[1] do quadrado do número de vértices (CORMEN et al., 2001).

[1] Apesar desta definição ser um pouco subjetiva, ela é comumente utilizada nos livros sobre o assunto.

Um grafo pode ser do tipo não-direcionado (também conhecido como não orientado) ou direcionado (orientado ou dígrafo) de acordo com o direcionamento das relações entre seus nós (DIESTEL, 2006). Em um grafo não direcionado, uma aresta é representada pelo par não ordenado de vértices u, v pertencentes ao conjunto V e indica que u e v possuem uma relação. Por exemplo, se o grafo representa uma rede social de amizades, a aresta u, v pode representar a amizade entre u e v. Neste caso, u é amigo de v e v é amigo de u. Já em um grafo direcionado uma aresta é representada por um par ordenado de vértices u, v pertencentes ao conjunto V e indica uma relação direcional de u para v. Por exemplo, em uma rede social de de seguidores de publicadores de conteúdo, como a rede formada a partir dos dados do *Twitter*[2], uma aresta u, v indica que o usuário u segue os comentários do usuário v. Esta aresta não indica nenhuma relação entre o usuário v e o usuário u, caso o usuário v siga o usuário u então haverá uma aresta v, u.

Uma aresta que conecta um nó a si mesmo é chamada de *laço* (ou auto-laço). Adicionalmente, é possível haver múltiplas arestas entre o mesmo par de vértices. Os grafos que possuem esta característica são chamados de *multigrafos*. Já os grafos que não possuem laços e nem arestas múltiplas são tipicamente chamados de *grafos simples* (DIESTEL, 2006). Os grafos que representam redes sociais no presente trabalho são grafos simples. É possível criar redes sociais representadas por multigrafos nos quais há diferentes tipos de arestas para representar diferentes relacionamentos (por exemplo, coautoria e orientação). Porém, nos estudos apresentados neste documento, optou-se por utilizar grafos simples e, quando necessário, foi criado um grafo para representar cada tipo de relacionamento em estudo.

Os grafos também podem ou não ser *ponderados*. Em grafos ponderados, cada aresta possui um peso associado a ela (CORMEN et al., 2001). Este peso pode indicar, por exemplo, a intensidade da relação entre os dois nós ou uma medida de distância ou custo entre eles.

Tipicamente, na análise de redes sociais, existe o foco no estudo de regiões específicas da rede. Um subgrafo de um grafo $G = (V, E)$ é um grafo G' cujo conjunto de vértices é um subconjunto de V' e seu conjunto de arestas é um subconjunto de E'. Obviamente as arestas pertencentes a E' devem ligar vértices pertencentes a V' (LEMIEUX; OUIMET, 2008).

Algumas medidas globais do grafo ou específicas de cada um de seus vértices são definidas e calculadas de forma diferente para grafos direcionados e não direcionados.

[2] https://twitter.com

Por simplificação, serão apresentadas apenas as definições relacionadas a grafos não direcionados, que são os mais utilizados no presente trabalho.

O *grau* de um nó corresponde ao número de arestas conectadas a ele (SZWARCFITER, 1986).

Um caminho (*path*) é um grafo $P = (V, E)$ tal que $V = (x_0, x_1, ..., x_k)$ e $E = (x_0x_1, x_1x_2, ..., x_{k-1}x_k)$, sendo $x_i \neq x_j$ para $i \neq j$. Diz-se que este caminho conecta o vértice x_0 ao vértice x_k e seu tamanho é igual ao número de arestas $|E|$, isto é, k (DIESTEL, 2006). O *caminho mínimo* (ou caminho geodésico) entre dois vértices em um grafo não ponderado é o caminho que possui o menor tamanho (número de arestas) entre todos os caminhos que conectam os dois vértices. O tamanho desse caminho, para grafos não ponderados, é também chamado de distância geodésica (WASSERMAN; FAUST, 2009).

Um *componente conexo* de um grafo é um subgrafo maximal no qual todos os nós estão conectados uns aos outros por caminhos. Se todos os nós de um grafo estiverem conectados uns aos outros dessa forma, o grafo será chamado de *conexo*. O componente conexo com maior quantidade de nós de um grafo é muitas vezes chamado de *componente gigante*. Diversas medidas utilizadas na análise de redes sociais só podem ser mensuradas em componentes conexos e, por isto, algumas análises realizadas neste trabalho consideraram apenas o componente gigante de cada rede (WASSERMAN; FAUST, 2009). A *porcentagem de nós no componente gigante* é uma medida bastante utilizada na análise de redes sociais. Pertencer ao componente gigante costuma ser associado a fazer parte do principal fluxo de conhecimento (ou informação) da rede. Assim, um valor elevado para esta medida costuma ser considerada uma característica positiva da rede social.

O *diâmetro de um grafo* é definido como a maior distância geodésica contida no grafo (NEWMAN, 2003b).

Um grafo simples é chamado de *grafo completo* se existirem arestas ligando cada um dos pares de vértices deste grafo. Assim, um grafo completo não direcionado possui $|V| * (|V - 1|)/2$ arestas. Um *clique* de um grafo é um subgrafo completo deste grafo (SZWARCFITER, 1986).

A *densidade* de um grafo corresponde à razão entre o número de arestas do grafo e o número máximo possível de arestas para o mesmo gráfico. Em um grafo simples não-direcionado, a densidade é calculada como $2 * |E|/(|V| * (|V - 1|))$ (SCOTT, 2009).

O *coeficiente de aglomeração* ou *coeficiente de clusterização* mede a transitividade das relações em um grafo. O coeficiente é calculado dividindo-se a quantidade de cliques

de tamanho três do grafo pela quantidade de trios de vértices que possuem ao menos um caminho de tamanho dois conectando-os. Em análise de redes sociais esta métrica muitas vezes é associada à estabilidade (ou maturidade) da rede (SCOTT, 2009; LEMIEUX; OUIMET, 2008).

Medidas de *assortatividade* verificam a tendência da existência de arestas entre vértices que compartilhem uma mesma características. Esta medida costuma variar de -1 (apenas vértices que não compartilham a característica são ligados) a 1 (todas as ligações ocorrem apenas entre vértices que possuem a característica). Diversas características podem ser consideradas, por exemplo, o grau do vértice (assim, a medida verificaria se há uma tendência de vértices com um mesmo grau estarem ligados). Para redes sociais acadêmicas, além da assortatividade de grau, também podem ser mensuradas a assortatividade de instituição (para verificar se pessoas de uma mesma instituição tendem a colaborar mais entre si do que com pessoas de outras instituições), assortatividade de estado, assortatividade de região do país, assortatividade de gênero, etc.

As medidas de *centralidade* visam a identificar o quão central (ou importante) um dado vértice é na rede de acordo com alguma característica. Existem diversas medidas de centralidade (FREEMAN, 1979; OTTE; ROUSSEAU, 2002), algumas das mais utilizadas são *centralidade de grau*, baseada no grau do vértice; *centralidade de proximidade (closeness)*, baseada na distância entre o vértice analisado e os demais da rede; *centralidade de intermediação (betweeness)*, baseada na frequência em que um dado vértice aparece entre todos os caminhos mínimos da rede; e *centralidade de autovalor (eigenvalue)*, baseada na importância dos vértices que estão ligados ao vértice em análise.

As medidas de *centralização* ou *centralidade da rede* são baseadas nas medidas de centralidade dos vértices da rede e servem para indicar o quão importante o vértice mais central de cada rede é para a sua rede (SCOTT, 2009). Centralizações altas em redes sociais costumam ser consideradas características negativas da rede, pois indicam que há um elemento muito central (influente ou importante), ou seja, a rede pode ser muito dependente deste elemento (FREEMAN, 1979; WASSERMAN; FAUST, 2009).

2.2 Plataforma Lattes e Currículo Lattes

A Plataforma Lattes é um sistema de informação que integra três importantes bases de dados: de Currículos, de Grupos de Pesquisa e de Instituições. Segundo o CNPq, esta

plataforma auxilia as ações de planejamento, gestão e operacionalização de fomento desta e de outras agências e fundações de apoio à ciência e tecnologia, bem como "para a formulação das políticas do Ministério de Ciência e Tecnologia e de outros órgãos governamentais da área de ciência, tecnologia e inovação" (Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), 2015).

Sobre o currículo Lattes, o CNPq faz as seguintes considerações:

"*O Currículo Lattes se tornou um padrão nacional no registro da vida pregressa e atual dos estudantes e pesquisadores do país, e é hoje adotado pela maioria das instituições de fomento, universidades e institutos de pesquisa do País. Por sua riqueza de informações e sua crescente confiabilidade e abrangência, se tornou elemento indispensável e compulsório à análise de mérito e competência dos pleitos de financiamentos na área de ciência e tecnologia.*" (Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), 2015).

Ao se realizar uma busca por currículos no *site* da Plataforma Lattes colocando apenas um espaço em branco no campo de busca, foram identificados 4.287.862 ao se consultar toda a base e 233.989 currículos de doutores (ativando a consulta apenas para doutores)[3]. Cada currículo Lattes está disponível em dois formatos (HTML e XML) e é dividido em oito seções principais: **dados gerais**, **formação**, **atuação**, **projetos**, **produções**, **eventos**, **orientações** e **bancas**. O conjunto de todos os currículos no formato XML ocupa cerca de 200GB de espaço em disco.

A seguir são destacadas algumas das informações presentes em cada uma das seções do Currículo Lattes. Nos *dados gerais*, encontram-se o nome completo do autor[4] do currículo, as formas citadas deste nome (preenchidas pelo possuidor), seu endereço profissional e a lista de prêmios e títulos.

Na *formação acadêmica/titulação*, encontram-se os dados referentes às diferentes formações/titulações, incluindo ensino fundamental, médio, profissional, curso de aperfeiçoamento, graduação, especialização, metrado e doutorado. Também nessa seção são cadastradas as informações referentes a pós-doutorados, livre-docência e formações complementares.

Os *projetos* são divididos em quatro grupos principais: pesquisa, desenvolvimento tecnológico, extensão, e outros tipos.

[3] Consultas realizadas no dia 10/08/2015 no site: http://buscatextual.cnpq.br/buscatextual/

[4] No presente trabalho, a pessoa a qual o currículo se refere será chamada de possuidor do currículo, autor, pesquisador ou doutor dependendo do contexto das análises realizadas.

As *produções* são dividas em três categorias: produção bibliográfica, produção técnica e produção artística/cultural.

Os *eventos* contêm as informações sobre participação em eventos, congressos, exposições e feiras. A seção *orientações* contém informações sobre sete tipos de orientação: dissertação de mestrado, tese de doutorado, monografia de conclusão de curso de aperfeiçoamento/especialização, trabalho de conclusão de curso de graduação, iniciação científica, supervisão de pós-doutorado, e orientação de outra natureza.

Por fim, nas *bancas* encontram-se informações tanto de participação em bancas de trabalhos de conclusão quanto da participação em bancas de comissões julgadoras

As informações presentes nos currículos da Plataforma Lattes são preenchidas manualmente pelos usuários. Existem alguns tipos simples de controle das informações (por exemplo, impedindo o cadastramento de informações sem o preenchimento de campos obrigatórios ou exigindo-se que algumas informações sejam obtidas de uma lista fornecida pela plataforma). Porém, não há nenhum tipo mais sofisticado de verificação da completude ou corretude da informação.

2.3 Análise bibliométrica

Existem diversos termos relacionados à análise quantitativa da produção científica, alguns dos mais usados são: “análise bibliométrica” ou “bibliometria”, “cientometria” e um termo que já está um pouco em desuso “bibliografia estatística” (*statistical bibliography*) (PRITCHARD, 1969). Há também outros dois termos relacionados: informetria e webometria. A seguir são dadas algumas definições desses termos.

Bibliometria (*bibliometrics*) é o estudo dos aspectos quantitativos da produção, disseminação e uso da informação. Na bibliometria são desenvolvidos modelos matemáticos e medidas para estes processos e então estes modelos e medidas são utilizados para a predição e tomada de decisão (TAGUE-SUTCLIFFE, 1992)

Cientometria (*scientometrics*) é o estudo dos aspectos quantitativos da ciências como uma disciplina. É uma subárea da ciência da informação, dentro das ciências sociais. Ela envolve estudos quantitativos das atividades científicas, incluindo, publicações e, assim, possui sobreposição com bibliometria (TAGUE-SUTCLIFFE, 1992). Cientometria corresponde a análise quantitativa da atividade de pesquisa científica, estudando tanto os recursos e os resultados quanto a organização e as técnicas de produção científica (CALLON et al., 1995)

Informetria (*informetrics*) é o estudo quantitativo dos aspectos da informação em todas as suas formas e em qualquer grupo social, incluindo os aspectos quantitativos da comunicação informal. Atingindo assim um contexto mais amplo do que da bibliometria e cientometria (TAGUE-SUTCLIFFE, 1992).

Webometria (*webometrics*) corresponde a análises informétricas do conteúdo disponível na web (ALMIND; INGWERSEN, 1997). Outra expressão correlata a webometria é a **cybermetria** (*cybermetrics*), que de um modo geral está sendo usada com o mesmo significado de webometria (VANTI, 2002).

Os trabalhos descritos no presente texto, enfocaram nas análises bibliométricas/cientométricas (SPINAK, 1998). As principais medidas utilizadas são as mesmas da maioria dos trabalhos na área e serão brevemente descritas a seguir.

Medidas baseadas no *número de artigos* publicados englobam a contabilização do número e artigos por pesquisador, por grupo de pesquisador, a média de artigos publicados por pesquisador ou pelo grupo e também variações de acordo com o tipo de publicação (em periódicos ou em anais, por exemplo). É importante destacar que, ao se contabilizar o total de publicações de um dado grupo, é necessário identificar as publicações em coautoria existentes dentro do grupo de forma a não contabilizar duas ou mais vezes o mesmo artigo. Este assunto será discutido na seção 2.6.

Medidas baseadas no *número de autores por artigo*. Englobam principalmente a contabilização e a média de autores por artigo (focando em um pesquisador ou num grupo de pesquisadores), podendo também contabilizar ou verificar a média de alguns tipos específicos de (co-)autores, por exemplo, quantidade média de orientados em cada artigo ou quantidade média de autores de um dado grupo por artigo.

Medidas relacionadas à *qualificação dos veículos de publicação*. São medidas que utilizam algum tipo de medida considerada qualitativa referente a revistas ou eventos (como JCR[5], SJR[6] ou Qualis[7]) para avaliar a produção. Vale ressaltar que estas medidas são feitas para qualificar revistas ou eventos, mas é comum encontrar artigos que aplicam essas medidas para qualificar a produção de pesquisadores individualmente ou de grupos de pesquisadores.

[5] http://wokinfo.com/products_tools/analytical/jcr/
[6] http://www.scimagojr.com/journalrank.php
[7] http://qualis.capes.gov.br/webqualis/

Medidas baseadas no *número de citações*. Englobam medições diretas da soma do número de citações dos artigos de um pesquisador ou grupo de pesquisadores e outras medidas ou índices derivados, como é o caso do índice h (HIRSCH, 2005) e do índice g (EGGHE, 2006). As citações são tipicamente computadas a partir de alguma base ou repositório que contenha citações, por exemplo, Scopus e Web of Science ou utilizando-se dados públicos extraídos do Google Scholar[8] e/ou do Microsoft Academic Search[9].

2.4 Obtenção, organização e refinamento dos dados

Os currículos da Plataforma Lattes são disponibilizados em dois formatos: HTML e XML. A versão XML é mais adequada para o processamento automático, pois possui todas as seções e campos dos currículos bem delimitados. Porém, na versão HTML, e em particular na versão exibida por um navegador de internet, há algumas informações não presentes na versão XML, por exemplo, o número de citações de cada artigo (das bases Web of Science[10], SciELO[11] e Scopus[12]) e se o pesquisador possui ou não bolsa produtividade e, em caso afirmativo, qual tipo e nível de bolsa ele possui. Nas pesquisas apresentadas neste documento, os dois tipos de arquivos foram utilizados.

Os currículos da Plataforma Lattes podem ser obtidos pela internet, utilizando-se, por exemplo, o comando *wget* ou a ferramenta *scriptLattes*. Porém, para isto, é necessário encontrar o identificar de cada currículo, utilizado para compor a URL (*Uniform Resource Locator*) completa do currículo.

A seguir são apresentadas algumas ferramentas ou trabalhos relacionados à obtenção e organização dos dados de Currículos Lattes.

Mena-Chalco e Cesar-Junior (2009) desenvolveram e disponibilizaram uma ferramenta de código aberto chamada *scriptLattes*[13]. Essa ferramenta recebe uma lista de identificadores de Currículos Lattes e produz como saída diversos arquivos no formato CSV (*Comma-separated value*) e páginas HTML organizando informações por categoria e também sumarizando informações por meio de gráficos e tabelas. Adicionalmente, a

[8] http://qualis.capes.gov.br/webqualis/
[9] academic.research.microsoft.com
[10] https://isiknowledge.com/
[11] www.scielo.org/
[12] www.scopus.com/
[13] http://scriptlattes.sourceforge.net/

ferramenta gera redes de colaboração, calcula algumas métricas sobre estas redes e utiliza a API do Google Maps para gerar um mapa com a distribuição geográfica dos pesquisadores. Essa ferramenta foi utilizada e/ou estendida por dezenas de trabalhos[14] que analisaram dados da produção científica nacional.

A figura 1 apresenta o componente gigante da rede de coautoria dos docentes do Bacharelado em Sistemas de Informação da EACH-USP produzida pela ferramenta *scriptLattes* utilizando dados de 2005 a 2014. A saída completa da ferramenta para este conjunto de dados de 41 professores pode ser encontrada em: `http://www.each.usp.br/digiampietri/si2014/`.

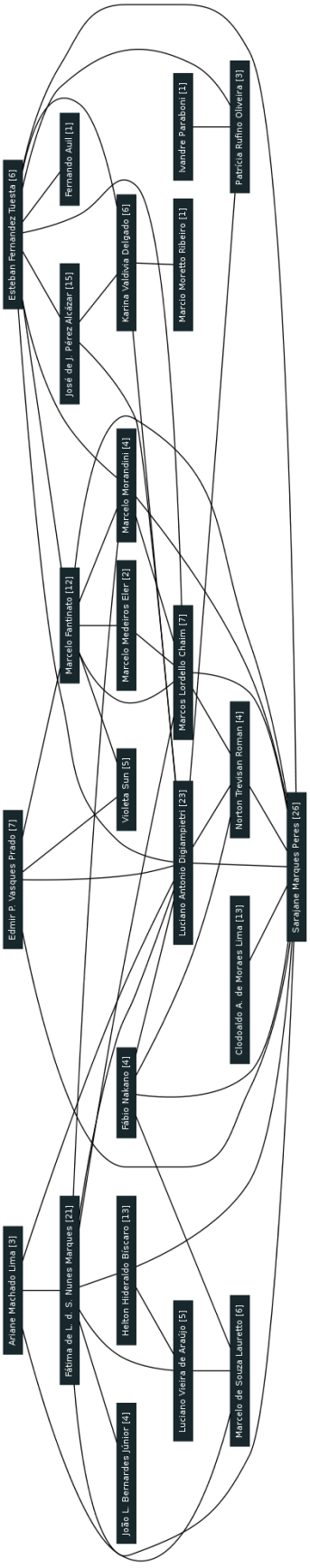
Alves, Yanasse e Soma (2011b) desenvolveram *LattesMiner*, uma linguagem multilingual específica de domínio para automatizar a extração de dados dos Currículos Lattes. Ela pode receber como entrada tanto o identificador do currículo como o nome do pesquisador/possuidor do currículo. *LattesMiner* foi utilizado para prover as entradas do sistema *Sucupira* (ALVES; YANASSE; SOMA, 2011a), um sistema para a visualização e análise de redes sociais acadêmicas. Dado um conjunto de currículos, o sistema produz a rede de coautorias e organiza em tabelas as informações sobre a produtividade dos pesquisadores.

Laender et al. (2011) apresentam parte do projeto *CiênciaBrasil*, que visa a prover ferramentas para auxiliar na organização e visualização da produção acadêmica brasileira. Entre as informações a serem organizadas está a rede de coautorias formada a partir de dados da Plataforma Lattes.

Na pesquisa descrita no presente documento, diferentes estratégias foram utilizadas para a identificação e organização de dados de currículos Lattes. Estas estratégias serão detalhadas no capítulo 3.

[14] Conforme pode ser observado, por exemplo, ao se analisar os resultados da busca por *scriptLattes* na ferramenta Google Scholar: https://scholar.google.com.br/scholar?q=scriptLattes

Figura 1 – Componente gigante da rede de coautoria dos docentes do Bacharelado em Sistemas de Informação da EACH-USP gerada pela ferramenta scriptLattes



Fonte: recorte da rede gerada pela ferramenta *scriptLattes*, imagem original disponível em: http://www.each.usp.br/digiampietri/si2014/grafoDeColaboracoes.html

2.4.1 Refinamento e enriquecimento dos dados

Existem diferentes atividades e abordagens envolvidas no refinamento e enriquecimento de um conjunto de dados. Uma das abordagens é a baseada na metodologia clássica de descoberta de conhecimento (GOLDSCHMIDT; PASSOS, 2005). Esta metodologia contém três etapas principais relacionadas a esse assunto:

- *Limpeza*: etapa na qual são tratados os problemas de dados com ruídos, incompletos ou inconsistentes. Por exemplo, há publicações cadastradas cujo título é composto apenas por um único espaço em branco.
- *Enriquecimento*: são agregadas informações aos dados, por exemplo, utilizando dados de outras fontes (como informações sobre os periódicos nos quais os artigos foram publicados ou o número de citações recebidas pelos artigos) para enriquecer os dados primários obtidos.
- *Construção de atributos*: nesta etapa são identificados quais atributos (características do conjunto de dados) podem ser compostos para formar novos atributos ou mesmo decompostos em atributos mais simples. É possível, nesta etapa, calcular os índices g e h de cada pesquisador (novos atributos) com base nas informações das citações de suas publicações.

As estratégias utilizadas neste projeto para refinamento e enriquecimento do conjunto de dados são descritas no capítulo 3.

2.5 Atualização, corretude e completude dos dados

Canibano e Bozeman (2009) afirmam em seu estudo sobre o uso de currículos *vitae* na política científica e avaliação da pesquisa que alguns dos principais fatores considerados na análise de dados, como a completude, corretude e atualização dos dados, são pouco tratados na literatura correlata.

Ao se tratar dos dados oriundos da Plataforma Lattes, estes fatores não são garantidos, pois os três dependem dos usuários que registram as informações que são disponibilizadas em seus currículos. Destes três fatores, a corretude tenta ser, minimamente, assegurada pela plataforma ao informar seus usuários sobre a responsabilidade legal das informações que são fornecidas. Já completude e a atualização dos dados não possuem

nenhum tipo de garantia ou mesmo um entendimento compartilhado do que elas significam dentro da plataforma. Por exemplo, não existe nenhuma recomendação sobre a frequência em que os currículos devem ser atualizados e mesmo que um currículo possua uma data de atualização recente isto não implica que todas as informações registradas no currículo estejam devidamente atualizadas.

Canibano e Bozeman (2009) destacam alguma características que sempre devem ser consideradas antes da utilização de qualquer fonte de dados, mas em especial no caso de fontes de currículos. Entre elas: (a) existência de algum mecanismo de validação das informações; (b) se os dados são preenchidos manualmente, o que pode levar a diversos problemas em seu tratamento, como falta de padronização e erros de digitação; (c) a frequência de atualização dos dados. Além disso, muitos campos dos formulários relacionados ao preenchimento dos currículos da Plataforma Lattes são opcionais, fato que pode limitar ou inviabilizar diversos tipos de análise (MARQUES, 2010).

Ainda sobre os dados da Plataforma Lattes, Silva e Smit (2009) destacam que essas características de cadastramento de informações na Plataforma Lattes podem comprometer a consistência dos dados dificultando sua recuperação ou a descoberta de informações a partir dos dados curriculares, o que pode limitar o uso desses dados.

Apesar de não haver garantias sobre a qualidade e atualização dos dados da Plataforma Lattes, a enorme quantidade de dados disponibilizados, a riqueza desses dados, bem como sua abrangência, tornam os dados dessa plataforma extremamente úteis para a análise da comunidade científica brasileira, justificando sua ampla utilização. Estes dados podem ser utilizados para diferentes estudos bibliométricos ou de análise de redes sociais acadêmicas, podendo servir de base para a elaboração de políticas científicas.

No capítulo 3 são descritas algumas das características dos dados da Plataforma Lattes utilizados nas pesquisas apresentadas neste documento, incluindo um estudo sobre a atualização dos currículos.

2.6 Resolução de entidades

A expressão "resolução de entidades" (*entity resolution*) corresponde ao processo de determinar se duas referência a objetos do mundo real se referem ou não ao mesmo objeto (TALBURT, 2010). Dentro do contexto de redes sociais acadêmicas, a resolução de entidades costuma ser utilizada principalmente para identificar se duas referências

a publicações se referem a mesma publicação, se duas referências a pessoas (autores, orientadores, etc) se referem a uma mesma pessoa, se duas referências a instituições se referem a uma mesma instituição, ou se duas referências a veículos de publicação correspondem ao mesmo veículo (periódico ou evento).

O processo de resolução de entidades pode ser dividido em cinco atividades principais (TALBURT, 2010):

1. Extração de referências a entidades: localização e coleta de referências de entidades, tipicamente a partir de informação não estruturada;
2. Preparação das referências das entidades: padronização, limpeza de dados e outras técnicas para aumentar a qualidade dos dados de referências;
3. Resolução das referências a entidades: decisão se duas referências se referem ou não à mesma entidade;
4. Gerenciamento das entidades identificadas: construção e manutenção de um registro das entidades resolvidas ao longo do tempo;
5. Análise do relacionamento das entidades: análise da rede de relacionamentos entre as diferentes entidades identificadas.

A atividade de resolução de nomes, propriamente dita, comumente utiliza técnicas que calculam a distância entre os nomes das referências a entidades (por exemplo, a distância de edição), além de utilizar de outras informações disponíveis. Por exemplo, no caso de autores e publicações, é possível utilizar modelos probabilísticos para avaliar a probabilidade de um dado artigo pertencer a um dado autor (com base no título do artigo, do veículo de publicação, do ano de publicação e/ou do assunto do artigo).

Ao problema específico de resolução de entidades aplicado a nomes de autores é dado o nome de desambiguação do nome do autor (*Author Name Disambiguation* - AND).

A desambiguação do nome do autor é fundamental para os trabalhos de bibliometria, pois pode evitar dois problemas: a atribuição incorreta de trabalhos a um dado autor no caso de autores homônimos ou a não atribuição de um trabalho a seu devido autor (por exemplo, quando o nome desse autor aparece de diferentes maneiras nas referências das publicações - polissemia). Destaca-se ainda que a homonímia é um problema bastante comum ao se analisar as publicações de trabalhos científicos, pois nestes trabalhos frequentemente não é apresentado o nome completo do autor e sim apenas seu primeiro e último nome ou a inicial do primeiro e o último nome, por exemplo.

Existem diversas bases de dados que tentam manter registros unívocos de autores, porém, a manutenção correta e atualizada desses registros é bastante custosa. Smalheiser e Torvik (2009) afirmam que só será possível confiar em um registro central de identificadores de autores (como o ORCID[15], por exemplo) se os autores participarem massivamente do processo de atualização e correção deste tipo de registro, tomando cuidado inclusive de alimentar essas bases com informações do passado. Porém, envolver um grande número de autores nesse processo é algo complexo, inclusive pelo fato da maioria dos autores possuir menos de três publicações e não terem motivação para manter uma base desse tipo correta e atualizada (SMALHEISER; TORVIK, 2009). Por outro lado, não há técnica de desambiguação automática de nomes que seja totalmente eficaz, as técnicas possuem acurácias diferentes e, tipicamente, possuem dificuldade em tratar nomes/sobrenomes comuns (como Lee e Smith). Estas características levaram muitos mantenedores de registros de identificadores e pesquisadores da área a trabalharem em duas direções diferentes: (a) desenvolvimento de algoritmos ou sistemas para a desambiguação automática de nomes que minimizem a ocorrência de falsos-negativos (mesmo que isso implique em uma taxa de revocação menor); ou (b) criação de sistemas semiautomáticos, que apenas sugerem que duas referências se referem ao mesmo autor e deixam para um ser humano decidir se a sugestão está correta ou não.

Milojevic (2013) afirma que mesmo as técnicas mais simples de desambiguação que consideram apenas a inicial do primeiro nome e o sobrenome conseguem obter resultados bastante satisfatórios (acurácia de até 97%), mas apenas quando usadas em alguns conjuntos de dados específicos. O autor explica que para bases com grandes volumes de dados e, especialmente, incluindo autores asiáticos, estas técnicas não funcionariam tão bem.

Strotmann e Zhao (2012) analisaram o impacto do uso de técnicas de desambiguação de nomes em análises bibliométricas. Os autores concluíram que as técnicas mais simples, porém ainda muito usadas, não apresentam resultados satisfatórios para diversas bases de autores e publicações. Em especial, apresentam o problema de multiautores (referências de múltiplos autores sendo agrupadas para um único autor) para bases de dados que possuem diversos autores com sobrenomes em chinês ou coreano. Apesar disso, os autores afirmam que as técnicas mais avançadas da área (que combinam diferentes atributos) possuem um

[15] http://orcid.org/

desempenho satisfatório e poderiam, assim, ser utilizadas para trabalhos de bibliometria sem comprometer de maneira significativa o resultado final desse tipo de trabalho.

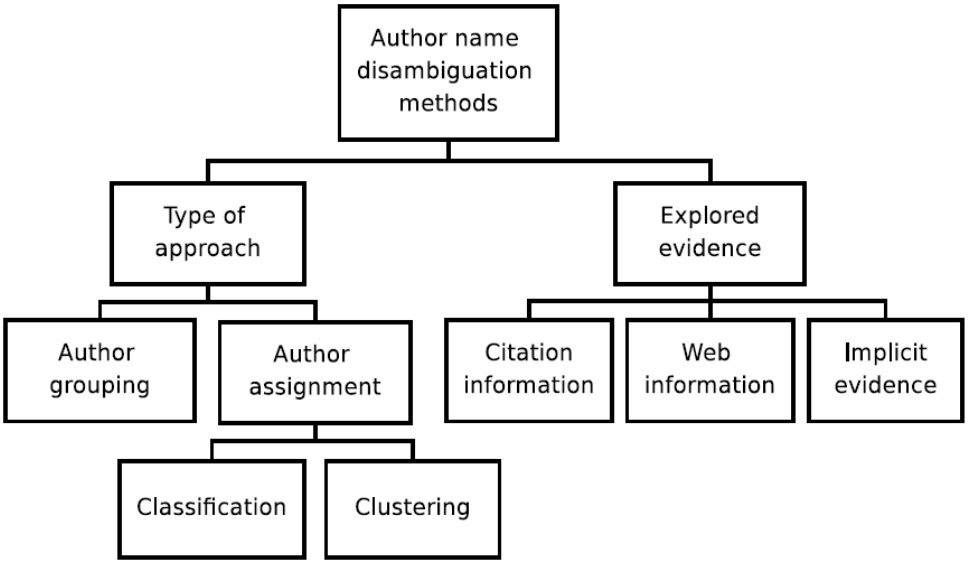
Um dos primeiros incrementos à utilização de apenas o nome do autor no processo de desambiguação é a análise de cocitação (WHITE; MCCAIN, 1988). Este tipo de análise parte da premissa que quanto maior a incidência de cocitação, maior a similaridade entre os citados. Algumas das técnicas consideradas complexas na desambiguação de nome de autores consistem da extração de diferentes atributos (ou características) dos dados disponíveis e do uso de uma estratégia para a tomada de decisão com base nesses atributos. Han et al. (2004) utilizaram duas estratégias, uma baseada em máquina de vetores de suporte e outra usando Naïve Bayes. Já Song et al. (2007) usam modelos Bayesianos Hierárquicos. Há também técnicas que utilizam lógica Fuzzy combinada com restrições passadas por um especialista do domínio para a desambiguação de nomes (DIAZ-VALENZUELA; MARTÍN-BAUTISTA; VILA, 2014).

Ferreira, Goncalves e Laender (2012), após revisarem algumas dezenas de trabalhos na área, propuseram uma taxonomia para descrever os métodos utilizados na desambiguação de nomes de autores (figura 2). A taxonomia descreve os métodos de acordo com duas características: a estratégia utilizada e as evidências utilizadas. Duas foram as estratégias identificadas pelos autores: o *agrupamento de autores*, no qual os dados são agrupados seguindo alguma função de similaridade e cada grupo deverá, idealmente, corresponder a todas as referências de uma entidade real (de um autor); e *atribuição direta de uma referência a um autor*, estratégia que, dada uma referência, atribuiu esta referência a um dado autor utilizando um modelo para representar cada autor (por meio de técnicas de classificação ou agrupamento, por exemplo).

Quanto às evidências utilizadas, Ferreira, Goncalves e Laender (2012) as classificaram em três: *informações da citação* como nomes dos autores, título do trabalho, veículo de publicação e ano de publicação; *informações oriundas da Web* correspondendo a informações adicionais mineradas da Web para enriquecer os dados das citações; e *evidências implícitas* que são informações inferidas a partir das demais informações (por exemplo, pode-se inferir o assunto de um artigo utilizando-se seu título e é possível utilizar um modelo probabilístico para modelar os assuntos publicados por um dado autor e, assim, averiguar a probabilidade de um título pertencer a um dado autor).

Além das diversas abordagens automáticas para a desambiguação de nomes de autores, há também estratégias semiautomáticas que usam a retroalimentação (*feedback*)

Figura 2 – Taxonomia para a classificação dos métodos de desambiguação do nome de autores



Fonte: Ferreira, Goncalves e Laender (2012)

fornecida pelos usuários para aprimorar os resultados de estratégias automáticas. Um exemplo desse tipo de abordagem foi desenvolvido por Ferreira, Machado e Goncalves (2012) e combina uma etapa de processamento não supervisionado com uma segunda etapa que envolve o *feedback* dos usuários. Os autores relatam que mesmo com um esforço pequeno (5% dos registros) de rotulação manual para resolver as ambiguações, foi possível obter uma melhora média de 10% no processo de desambiguação. Uma estratégia semelhante foi utilizada por Godoi et al. (2013), que afirmam ter obtido melhores resultados do que os algoritmos do estado-da-arte para desambiguação de nomes de autores que não utilizam retroalimentação fornecida pelos usuários.

No projeto de pesquisa apresentado neste trabalho, diferentes estratégias de resolução de entidades formam utilizadas conforme será detalhado no capítulo 4.

2.7 Análise de grupos

Conforme apresentado, a análise de redes sociais visa a estudar características da interação entre diferentes indivíduos dentro de uma rede. A análise de redes sociais acadêmicas, pelo fato de muitas vezes combinar a análise de redes sociais com aspectos da bibliometria, tipicamente aborda a análise de grupos em duas vertentes: análises bibliométricas e análises de redes sociais.

Os grupos a serem analisados podem ser formados de diferentes maneiras. Por exemplo, de acordo com a formação/titulação dos indivíduos, de acordo com sua localização geográfica (país, estado ou cidade), de acordo com suas áreas de atuação ou com o local/instituição em que trabalha (universidade, departamento ou programa de pós-graduação)

A avaliação da produtividade (muitas vezes relacionada a análises bibliométricas) de grupos de pesquisa tem se tornado cada vez mais relevante, uma vez que há uma quantidade limitada de recursos para fomentar a pesquisa e um número cada vez maior de pesquisadores ou instituições interessados nesses recursos. Quanto mais abrangente e correta for a análise, maior será a possibilidade de se alocar os recursos de maneira meritocrática. Além disso, o conhecimento das características dos grupos de pesquisa de um dado estado, região ou do país é fundamental para a elaboração de políticas científicas eficazes. Porém, este tipo de avaliação é uma tarefa extremamente complexa, pois envolve a análise de diferentes características tanto quantitativas como qualitativas, muitas das quais possuem natureza subjetiva (DIGIAMPIETRI et al., 2014). Além disso, não existe um consenso sobre quais métricas ou características devem ser consideradas e quais pesos devem ser atribuídos a cada uma.

Na avaliação de grupos acadêmicos e, em especial, focando-se em departamentos ou programas de pós-graduação, há algumas métricas comumente utilizadas (LAENDER et al., 2008). As cinco métricas a seguir estão presentes nos documentos de área da CAPES[16] (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) e são utilizadas para a avaliação dos programas brasileiros de pós-graduação: objetivos do programa; corpo docente; estudantes; produção intelectual; e inserção social.

Ao se tratar da avaliação de redes sociais acadêmicas considerando características geográficas (cujos elementos são agrupados de acordo com o estado ou o país), há alguns aspectos adicionais que precisam ser tratados, como a coleta de grandes volumes de dados. Ao se realizar uma análise da rede acadêmica brasileira é importante lembrar que o Brasil é o quinto maior e quinto mais populoso país do mundo. Isto combinado à grande diversidade cultural, geográfica e social torna a análise da comunidade científica brasileira ainda mais interessante e desafiadora.

A colaboração científica é influenciada por diversos fatores como proximidade geográfica e os relacionamentos anteriores dos pesquisadores (como a relação orientador-

[16] http://www.capes.gov.br/

orientado) (NEWMAN, 2004; GLANZEL; SCHUBERT, 2004; MENEZES et al., 2009). A informação necessária para o estabelecimento das redes sociais acadêmicas e avaliação de grupos de pesquisa é tipicamente de diferentes naturezas e dispersa. Muitos desses diferentes tipos de informação são agrupados nos currículos dos pesquisadores, o que ajuda a enfrentar alguns dos desafios encontrados especialmente ao se utilizar apenas bases de citações bibliográficas, como: homonímia e polissemia, a dispersão da produção bibliográfica em diferentes fontes, a afiliação do pesquisador, as relações de orientação, áreas de interesse, identificação das colaborações/coautorias, etc.

Ao se pensar na avaliação de toda a rede social acadêmica de um país qualquer, o primeiro obstáculo (e muitas vezes um dos maiores) é a obtenção do grande volume de informação relacionada, normalmente distribuída em dezenas ou centenas de fontes de informação (como páginas Web e relatórios de universidades, do governo, bibliotecas digitais internacionais, etc). No entanto, esta atividade é bastante simplificada no Brasil, pela existência da Plataforma Lattes (descrita na seção 2.2).

Nos últimos anos, diversos trabalhos analisaram a produtividade de grupos de pesquisa, alguns com enfoque bibliométrico (HIRSCH, 2005; BOLLEN; RODRIQUEZ; SOMPEL, 2006; GARFIELD, 1955; DUFFY et al., 2011; MARTINS et al., 2010) e outros combinando métricas da análise de redes sociais (MENEZES et al., 2009; MENA-CHALCO; CESAR-JUNIOR, 2009; FRANCESCHET, 2011).

Algumas das principais medidas bibliométricas utilizadas foram o fator de impacto, índice-h e número de citações (GARFIELD, 1955; HIRSCH, 2005). Sendo que alguns trabalhos ponderam a participação dos autores nas publicações de acordo com a posição do pesquisador na lista de autorias, por exemplo, o primeiro autor recebe mais peso (DUFFY et al., 2011). Além da análise de grupos de pesquisadores de uma dada região ou área do conhecimento, alguns destes trabalhos também analisam se há influência/correlação entre gênero e as demais medidas além de analisar a evolução na carreira de alguns pesquisadores.

Na área da Ciência da Computação, há uma discussão importante sobre como computar artigos publicados em anais de eventos na análise bibliométrica (VARDI, 2009) (na maioria das outras áreas são considerados apenas os artigos publicados em revistas). Martins et al. (2010) avaliaram as conferências de acordo com a quantidade de citações recebidas pelos seus artigos e destacaram que são necessários novos mecanismos para avaliar as conferências. Ainda sobre o uso de indicadores para avaliar grupos, Franceschet (2010) realizou um estudo comparando os mesmos indicadores, porém obtidos de fontes

diferentes (Web of Science e Google Scholar) e concluiu que há uma grande correlação entre as medidas com algumas variações especialmente ao se comparar índices como o índice-h.

Menezes et al. (2009) combinaram a análise bibliométrica com a análise de redes sociais para analisar a área de Ciência da Computação e suas subáreas em diferentes regiões do mundo (Estados Unidos, Europa e Brasil) no período de 1994 a 2006. Nesse trabalho, os autores destacaram os aspectos comuns e, principalmente, as diferenças observadas entre as redes de cada região.

Franceschet (2011) utilizou dados do DBLP[17] (LEY, 2002) a fim de analisar as redes de coautorias na área de Ciência da Computação e compará-las com dados de redes de outras áreas. Ele observou que, em comparação com outras áreas, o nível de colaboração (em termos de números de coautorias nas publicações) em Ciência da Computação é baixo ou moderado. O autor também constatou que relacionamentos mais fortes (isto é, reincidentes e mais duradouros) costumam ser mais frequentes nas colaborações em publicações em artigos de revistas do que nos artigos de conferências.

Considerando a análise da produtividade da rede acadêmica de Ciência da Computação brasileira (estudo realizado nesta pesquisa que será detalhado no capítulo 5), Laender et al. (2008) compararam alguns indicadores do programas brasileiros considerados de nível internacional com alguns programas da América do Norte e da Europa e concluíram que os programas brasileiros analisados atingiram a maturidade. Os autores partiram de dados extraídos do DBLP.

Também utilizando dados do DBLP, Freire e Figueiredo (2011) analisaram a rede social acadêmica brasileira na área de Ciência da Computação. Em particular, os autores detectaram a presença de *super peers*, isto é, alguns indivíduos com grau muito acima dos demais da rede. Os autores também propuseram uma métrica para a avaliação de redes de colaboração que considera a importância de um nó na conexão de indivíduos de diferentes grupos.

Informações da Plataforma Lattes vêm sendo cada vez mais utilizadas em análises de grupos. A maioria destas análises estuda grupos pequenos (com poucas dezenas de pesquisadores) e muitas vezes a organização da informação utilizada é realizada de maneira manual ou semiautomática. Três exemplos de estudos que utilizaram dezenas de milhares

[17] http://dblp.uni-trier.de/db/

de currículos Lattes são os trabalhos de Mena-Chalco e Cesar-Junior. (2011), Medeiros e Mena-Chalco (2013) e Melo (2011).

Mena-Chalco e Cesar-Junior. (2011) estudaram a rede de coautorias utilizando os dados dos doutores que atuam em uma das quatro seguintes áreas: Ciência da Computação, Matemática, Física e Estatística. Os autores focaram a análise dos grupos considerando a distribuição dos graus dos nós e dos valores do *AuthorRank* (LIU et al., 2005).

Em sua tese de doutorado, Melo (2011) analisou o currículo de mais de 51 mil pesquisadores que são participantes de grupos de pesquisa. A autora objetivou caracterizar a comunidade científica brasileira considerando três aspectos: produtividade, internacionalização e visibilidade.

Medeiros e Mena-Chalco (2013) analisaram mais de 650 mil currículos da Plataforma Lattes a fim de estudar a rede social composta por todas as pessoas que declararam atuar em ao menos uma das seguintes grandes-áreas: Ciências Humanas, Ciências Sociais Aplicadas ou Linguística, Letras e Artes. Os autores mediram, ao longo dos anos, o comportamento de alguns aspectos bibliométricos e algumas métricas de redes sociais para o conjunto de dados analisado encontram algumas características da dinâmica das redes formadas. Adicionalmente, realizaram uma análise de frequência das palavras nos títulos das publicações para identificar quais palavras estão sendo mais utilizadas e que, de certa forma, são as mais importantes para estas áreas.

Dentro do projeto de pesquisa ao qual o presente texto se refere, diferentes estudos de grupos foram realizados. Três tipos se destacam: a análise de "toda" a rede brasileira de pesquisadores (MENA-CHALCO et al., 2014), estudos mais aprofundados da rede formada pelos docentes dos programas de pós-graduação em Ciência da Computação no Brasil (DIGIAMPIETRI et al., 2012b; DIGIAMPIETRI et al., 2014; DIGIAMPIETRI et al., 2015b) e análises das redes dos doutores que atuam no Brasil agrupados pelo estado no qual trabalham (DIGIAMPIETRI et al., 2014a). Um detalhamento sobre os dois últimos tipos de estudo é apresentado no capítulo 5.

2.8 Análise de tendências

Existem diversas definições para o termo tendência, sem haver uma definição comumente aceita nas diferentes áreas. Porém, uma característica frequentemente encontrada nas descrições da palavra tendência é a propensão de um dado objeto (ou de um de seus

atributos) em realizar algum comportamento (ou uma mudança de valor). Por exemplo, pode-se verificar que a ação de uma dada empresa apresenta tendência de alta (isto é, projeta-se que o valor dessa ação irá aumentar). Outro fator intrinsecamente ligado à análise de tendências é o fator temporal: a tendência é a identificação de um comportamento ou propensão em um dado período de tempo.

Vejlgaard (2008) apresenta algumas das definições para o termo análise de tendências em especial dentro da área de sociologia. Segundo o autor, analisar tendências é o processo de observar as mudanças no comportamento de indivíduos ou grupos. Assim, tendências correspondem a padrões de comportamento social ou estilo de vida observados ao longo do tempo (VEJLGAARD, 2008).

Já para Han, Kamber e Pei (2006), análise de tendências consiste no processo de modelar um conjunto de dados utilizando séries temporais de forma a entender o comportamento desses dados e prever valores futuros. Este tipo de definição é bastante utilizado quando os dados trabalhos são números, os quais possuem algum rótulo temporal (por exemplo, para análise de tendências dentro do mercado de ações no qual os dados são os valores de uma dada ação ao longo do tempo).

A análise de tendências também pode ser aplicada a documentos textuais (dentro da Mineração de Textos). Kontostathis, Galitsky e Pottenger (2004) definem uma tendência como um tópico que cresce em interesse ou utilidade ao longo do tempo. Um dos desafios da análise de tendências aplicadas a documentos textuais é a identificação de tópicos (das unidades de informação que se pretende verificar a tendência). A identificação destes tópicos é conhecida como *Emerging Trending Detection (ETD)* (BERRY, 2013).

Outro aspecto que pode ser considerado durante a análise de tendência é o conjunto de características das fontes geradoras de conteúdo (GLOOR et al., 2009). Isto é, além de se considerar o conteúdo propriamente dito, é possível também considerar características individuais de quem gerou este conteúdo, bem como a situação desta fonte (que pode ser uma pessoal) dentro de sua rede de atuação (por exemplo, um pesquisador produzindo publicações e que se encontra dentro de uma rede social acadêmica).

A análise de tendências pode ser aplicada em diferentes contextos. Dentro do contexto acadêmico, muitas vezes tenta-se identificar quais assuntos estão ganhando destaque nos últimos anos. Esta identificação pode ser feita observando-se, por exemplo, a quantidade de eventos que surgem em uma dada (sub)área, a quantidade de artigos publicados num dado assunto e/ou a quantidade de citações que artigos de uma área

ou assunto estão recebendo. A identificação se um artigo pertence a uma área ou a um assunto, por si só, é uma tarefa complexa (conforme será discutido na próxima seção) e pode ser feita de diferentes maneiras de acordo com as informações disponíveis. As mais comuns são: título do artigo, seguido pelo resumo e palavras chaves, mas há trabalhos que utilizam todo o conteúdo do artigo para a identificação de seu assunto e/ou da área em que está contextualizado. Rowley (1994) discute algumas vantagens e desvantagens da utilização de pequenos recortes de informação (como o título) na indexação e recuperação de informações.

A análise de tendências dentro do contexto acadêmico pode ser utilizada, por exemplo, para a elaboração de políticas científicas focando em assuntos ou temas em ascensão, como guia para agências de fomento sobre o contexto atual da pesquisa proposta em um projeto ou para auxiliar pesquisadores a verificar se o comportamento recente de um tema ou assunto (para, por exemplo, auxiliar o pesquisador que tem interesse em estudar um novo assunto).

Um detalhamento sobre métodos, técnicas e aplicações da análise de tendências e, em especial, aqueles que utilizam características das fontes geradoras de informação ou ao menos possuem aplicações técnico-sociais, pode ser encontrado na revisão sistemática realizada sobre o assunto no segundo semestre de 2013 (TRUCOLO; DIGIAMPIETRI, 2014b).

Na seção 7.2 são apresentados alguns dos resultados obtidos com a especificação e implementação de uma técnica para a análise de tendências que combina técnicas tradicionais da análise de tendências usando dados textuais com o uso de métricas oriundas da análise de redes sociais.

2.9 Identificação de áreas do conhecimento

A identificação das áreas de atuação de um dado pesquisador e/ou áreas ou assuntos de um artigo científico é uma atividade importante para a identificação de especialistas ou criação de grupos temáticos para posterior análise.

Há diferentes abordagens para esta identificação que variam bastante de acordo com as informações disponíveis. Dentre as informações curriculares tipicamente utilizadas destacam-se: informações sobre as publicações (títulos de artigos, palavras-chave, categorias, veículo de publicação, resumo do artigo ou mesmo o artigo completo), informações sobre projetos de pesquisa, departamento ou instituição na qual o pesquisador trabalha e

disciplinas ministradas. Adicionalmente, informações sobre a rede social acadêmica do pesquisador também podem ser utilizadas, especialmente para os casos em que já se sabe de antemão as áreas de atuação de alguns dos colaboradores do pesquisador. Dois exemplos de medidas que podem ser utilizadas considerando a rede social acadêmica são: porcentagem dos vizinhos na rede que atuam em cada área e porcentagem dos vizinhos ou vizinhos dos vizinhos (vizinhança nível 2) que atuam em cada área.

A identificação de áreas costuma ser tratada dentro da Análise de Redes Sociais por meio de algoritmos de inteligência artificial, modelos de Markov ou análise de vizinhança (WANG; KRIM, 2012; WANG; KRIM; VINIOTIS, 2013). Por outro lado, o uso de textos livres (por exemplo, títulos ou resumos de artigos) para identificação de categorias ou extração de conhecimento é tipicamente tratado por técnicas de mineração de textos ou processamento de língua natural (GHAREHCHOPOGH; KHALIFELU, 2011; GERDSRI; KONGTHON; PUENGRUSME, 2012).

Na seção 7.1 são apresentados resultados combinando técnicas de mineração de textos e análise de redes sociais para a identificação automática das grandes-áreas, áreas e subáreas de atuação de pesquisadores com base no títulos de suas publicações e de suas redes de coautoria.

2.10 Recomendação de conteúdo

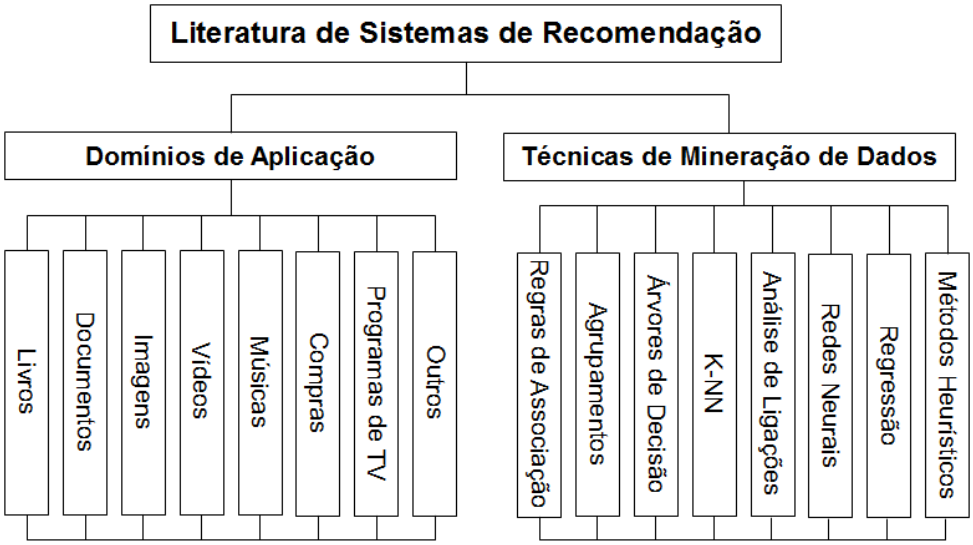
Sistemas de recomendação de conteúdo visam a auxiliar no processo de tomada de decisão, especialmente nos casos em que uma pessoa não possui experiência pessoal suficiente sobre todas as alternativas (RESNICK; VARIAN, 1997). Estes sistemas fornecem sugestões de "itens" a serem selecionados por um usuário, por exemplo: quais itens comprar, quais músicas ouvir ou quais notícias ler (RICCI et al., 2011; CAZELLA; NUNES; REATEGUI, 10; LU et al., 2012). A recomendação também pode ser de pessoas: possíveis amigos numa rede social online, pessoas a serem seguidas num sistema de blog/microblog, pessoas com quem trabalhar num dado projeto, especialistas em um dado assunto, etc (REATEGUI; CAZELLA, 2005).

A necessidade da recomendação personalizada de conteúdo está cada vez maior, pois há uma crescente quantidade de informações/itens disponíveis e objetiva-se minimizar o tempo necessário despendido para se encontrar um item relevante. Considerando-se redes sociais acadêmicas, a recomendação de conteúdo pode ser aplicada para diversos fins,

entre eles: artigo ou livro a ser lido; potenciais colaboradores para trabalhos científicos; especialista em um dado domínio (para solicitação de um parecer ou participação de uma banca, por exemplo).

A tabela 3 apresenta uma estruturação dos sistemas de recomendação organizada de acordo com as principais aplicações encontradas na literatura e as principais técnicas utilizadas (PARK et al., 2012).

Figura 3 – Estruturação dos sistemas de recomendação de conteúdo.



Fonte: Brito e Digiampietri (2013), adaptado de Park et al. (2012)

De um modo geral, pode-se classificar os sistemas de recomendação em três categorias principais (REATEGUI; CAZELLA, 2005): (i) Baseados em Conteúdo (*Content-Based Filtering*): sistemas que recomendam itens ao usuário com base nas características do item e informações sobre o usuário (por exemplo, recomenda-se itens similares aos itens que o usuário já selecionou); (ii) Filtragem Colaborativo (*Collaborative Filtering*): recomendações baseadas nas avaliações dos itens realizadas por um conjunto de usuários cujos perfis são mais similares ao do usuário alvo (este conjunto de usuários pode ser formado, por exemplo, por usuários com avaliações semelhantes às do usuário alvo, vizinhos numa rede social ou geograficamente próximos); (iii) Sistemas Híbridos: que combinam as características das duas categorias anteriores.

Os sistemas Baseados em Conteúdo e que utilizam Filtragem Colaborativa apresentam algumas características complementares e por isso os Sistemas Híbridos vem se destacado atualmente por reduzirem as limitações dos dois primeiros. Em particular, os sistemas apenas Baseados em Conteúdo têm a vantagem de que um novo item não

precisa ser previamente avaliado por outros usuários para poder ser recomendado (pois a recomendação será baseada no casamento entre as características do item e o perfil de seleção de itens do usuário). Por outro lado, este tipo de sistema pode gerar o problema de superespecialização, isto é, só serão recomendados itens muito parecidos com os itens já selecionados e/ou recomendados ao usuário. Já os sistemas que só utilizam a Filtragem Colaborativa apresentam o problema do novo item: este ainda não foi selecionado/recomendado por nenhum usuário então não será considerado para as recomendações. Porém, este tipo de sistema apresenta a vantagem da característica serendipista, isto é, ele poderá fazer recomendações surpreendentes (no sentido de bem diferentes dos itens já selecionados) e relevantes aos usuários ao passo que as sugestões serão feitas com base nos itens selecionados por usuários com um perfil semelhante.

Além das informações sobre os itens a serem recomendados, informações do perfil do usuário e informações da rede social do usuário, algumas outras informações também estão sendo utilizadas por sistemas de recomendação: informação de contexto, *folksonomies*, ontologias ou taxonomias.

Dentre as informações de contexto mais utilizadas estão o clima, a época do ano e o local onde o usuário se encontra (JULASHOKRI; FATHIAN; GHOLAMIAN, 2010; RATTANAJITBANJONG; MANEEROJ, 2009). Estas informações são utilizadas principalmente em sistemas de recomendação de viagens ou de montagem de itinerários para turísticas (ALABASTRO et al., 2010; IAQUINTA et al., 2009), ou para recomendar notícias (YEUNG; YANG, 2010).

Outra informação utilizada é a folksonomia (*folksonomy* ou *social/collaborative tagging*), que consiste de informações inseridas por usuários para descrever um dado item. Em particular, este tipo de informação é bastante utilizada para a recomendação de conteúdos não textuais, como obras de arte (LOPS et al., 2009), filmes e programas de TV (BARRAGANS-MARTINEZ et al., 2010).

Ontologias ou taxonomias são utilizadas para estabelecer relacionamentos entre itens que não seriam possíveis de identificar apenas pela descrição dos itens. Apesar do grande poder semântico que pode ser representado em uma ontologia, a maioria dos estudos utiliza apenas uma taxonomia para organizar diretamente os itens ou as palavras-chaves que os descrevem. Mesmo assim, este tipo de informação contribui para a recomendação (WAN et al., 2010; LOH et al., 2008). Além da recomendação de produtos, este tipo de informação vem sendo usada para a recomendação de notícias (CANTADOR; BELLOGIN; CASTELLS, 2008) e artigos científicos (PUDHIYAVEETIL et al., 2009).

Diversas extensões à Filtragem Colaborativa surgiram nos últimos anos (BERNARDES et al., 2015) e as principal diferença entre elas está na formação do grupo dos usuários que são considerados os vizinhos do usuário alvo e/ou na ponderação dada a cada vizinho. Na abordagem mais tradicional, vizinhos são aqueles que apresentam um perfil de seleção de itens mais semelhante ao perfil do usuário alvo. Em abordagem mais recentes, conhecidas como Recomendação Social, os vizinhos podem ser seus amigos de uma rede social online (RICCI; ROKACH; SHAPIRA, 2011), ou pode-se criar redes de confiança com base em algumas métricas de redes sociais (JAMALI; ESTER, 2009; MA et al., 2011; YANG; STECK; LIU, 2012; MEYFFRET; MÉDINI; LAFOREST, 2012; GAO; XU; CAI, 2011) ou mesmo por meio de algum tipo de indicação direta do usuário (JAMALI; ESTER, 2010).

Dentro do presente projeto de pesquisa investiga-se como utilizar informações da análise de redes sociais para auxiliar na atividade de recomendação de conteúdo, em particular, recomendação de leitura de artigos científicos. Este aspecto da pesquisa ainda está em etapa inicial de desenvolvimento.

2.11 Dinâmica da rede

A dinâmica de uma rede social costuma ser analisada principalmente de acordo com o surgimento ou exclusão de arestas entres os indivíduos da rede, mas há outras características que também podem ser consideradas. Dentre elas, estão o aparecimento ou a exclusão de indivíduos (nós), a variação de atributos relacionados às arestas (por exemplo, variação no peso das arestas), e a variação nos atributos estruturais da rede (AGGARWAL; SUBBIAN, 2014).

A análise do surgimento de novos relacionamentos (arestas) entre os indivíduos é estudada pela predição de relacionamentos, que estuda o comportamento dos relacionamentos ao longo do tempo para tentar predizer novos relacionamentos. Este tipo de predição pode ser empregada de diferentes formas: predição de relacionamentos inéditos entre indivíduos; predição da reincidência de relacionamentos entre indivíduos; e até mesmo a predição do fim do relacionamento entre indivíduos. Na próxima seção será apresentada uma descrição sobre a área de predição de relacionamentos em redes sociais.

Diversos trabalhos que envolvem a análise de redes sociais avaliam aspectos da dinâmica das redes. A maioria dos estudos é focada em grupos específicos, como autores

que publicaram artigos sobre um determinado tema (HORN et al., 2004; SHARMA; URS, 2008) ou que publicaram artigos em um evento científico específico (HAYAT; LYONS, 2010).

Outros trabalhos utilizam a dinâmica das redes sociais online para identificar padrões temporais na produção de conteúdos. Guo et al. (2009) analisam a produção e divulgação de conteúdos em diferentes tipos de redes sociais online. Os autores conseguiram identificar diferentes padrões temporais na produção, qualidade e esforço relacionados à produção dos conteúdos.

Há trabalhos focados no desenvolvimento de modelos ou ferramentas para facilitar a visualização e a análise da dinâmica das redes sociais (BERGER-WOLF; SAIA, 2006; KANG; GETOOR; SINGH, 2007), incluindo sistemas para simulação do comportamento de redes (BAUMES et al., 2008).

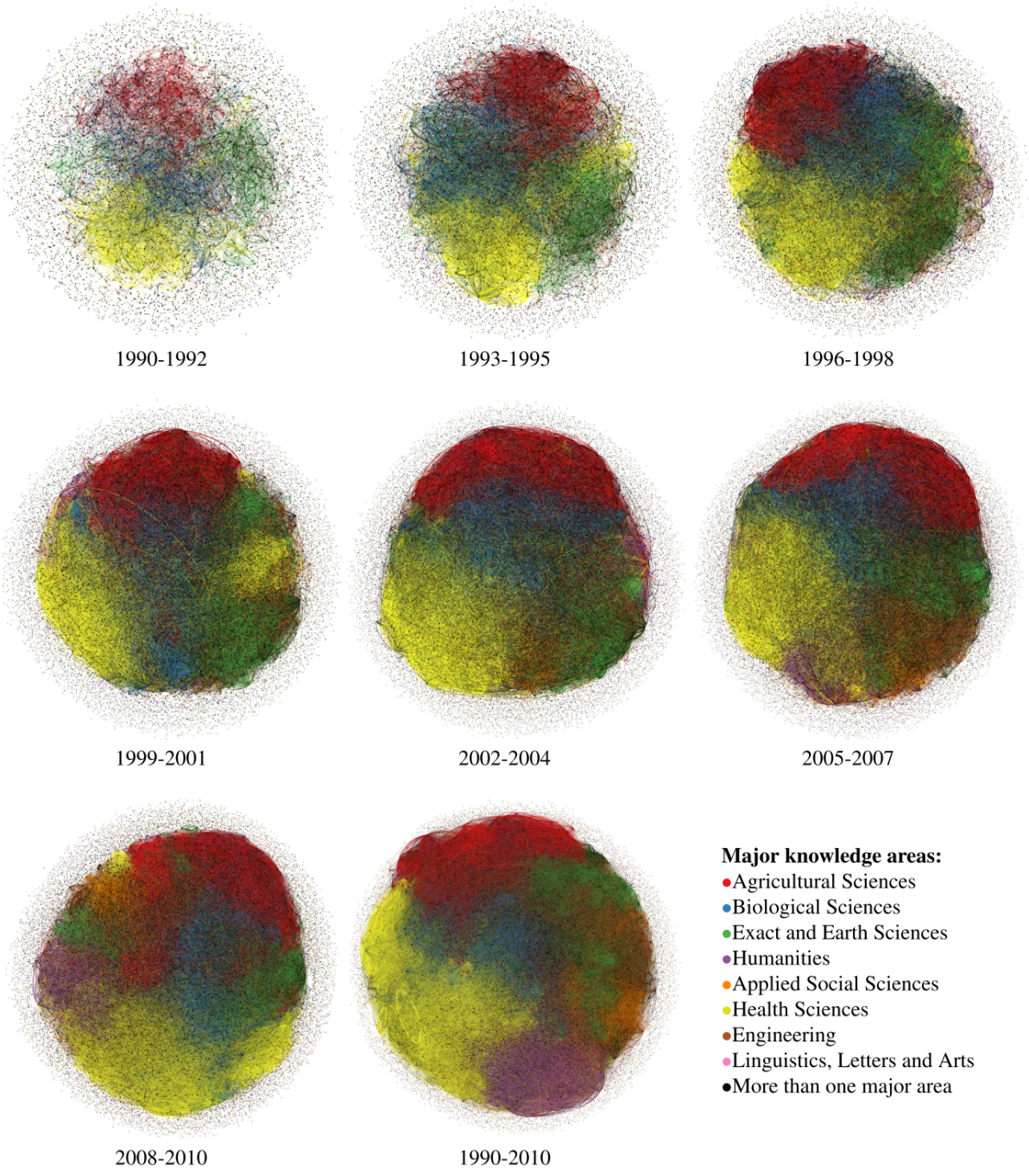
A figura 4 ilustra a evolução da rede de coautorias brasileira, considerando dados extraídos de mais de um milhão de currículos da Plataforma Lattes (MENA-CHALCO et al., 2014). Os nós e as arestas foram coloridos de acordo com as diferentes grandes-áreas de atuação cadastradas nos currículos dos pesquisadores.

Aspectos relacionados à dinâmica de redes sociais acadêmicas foram estudados neste projeto de pesquisa nos seguintes trabalhos Digiampietri et al. (2012b), Digiampietri, Santiago e Alves (2013), Digiampietri et al. (2014), Mena-Chalco et al. (2014), Digiampietri et al. (2015b) e Digiampietri et al. (2015).

2.12 Predição de relacionamentos

Um dos aspectos mais pesquisados em relação à dinâmica de redes sociais é o surgimento de novos relacionamentos. Uma das áreas de estudo que ganhou destaque nos últimos anos é a predição de relacionamentos (*link prediction*), pois pode ser utilizada tanto para encontrar amigos que ainda não estavam ligados em numa rede social online (BARBIERI; BONCHI; MANCO, 2014; VASUKI et al., 2010; TIAN et al., 2010; PEREZ; BIRREGAH; LEMERCIER, 2012; FIRE et al., 2011; ZHONG et al., 2013; QUERCIA; CAPRA, 2009) quanto para potencializar a realização de trabalhos em uma comunidade científica ou em empresas (HSIEH et al., 2013; DONG et al., 2012; SA; PRUDENCIO, 2011). Porém, o comportamento fortemente dinâmico de uma rede social torna a predição de relacionamentos uma atividade complexa (LIBEN-NOWELL; KLEINBERG, 2003; LU; ZHOU, 2011).

Figura 4 – Evolução da rede de coautorias brasileira



Fonte: Mena-Chalco et al. (2014)

Nas redes sociais acadêmicas, a predição de relacionamentos tem sido utilizada principalmente para a predição de coautorias (DONG et al., 2011; GAO; DENOYER; GALLINARI, 2012; GUO; GUO, 2010; LIN; YUN; ZHU, 2012; MAKREHCHI, 2011). Isto é, predizer se um par de pesquisadores irá colaborar na publicação de um artigo. Este tipo de predição pode ser utilizada para, além de estimar o comportamento futuro de uma rede social acadêmica, auxiliar na indicação de parcerias potencialmente promissoras.

A predição de coautorias comumente aborda a predição de coautorias novas/inéditas (isto é, visa a prever quais pares de autores que nunca colaboraram na publicação de um artigo irão colaborar). Porém, também é possível realizar a predição de uma forma mais geral: prever quais pares de autores irão colaborar na publicação de um artigo independentemente deles já terem ou não colaborado previamente.

Para a realização da predição de coautorias os seguintes fatores costumam ser considerados (HASAN; ZAKI, 2011): identificação ou especificação de atributos/métricas a serem utilizados (estes atributos pode ser relacionados ao perfil de cada autor ou podem ser métricas estruturais oriundas da análise de redes sociais); definição da estratégia para agregar ou combinar os diferentes atributos; e estabelecimento da estratégia que será utilizada para tratar os dados desbalanceados (isto é, dado um par arbitrário de autores, é mais provável que eles não irão colaborar na publicação de um artigo).

Nos últimos anos, diversos trabalhos propuseram técnicas para a predição de relacionamentos (HASAN; ZAKI, 2011; IBRAHIM; CHEN, 2015; WANG et al., 2015). Dentre estes trabalhos, há aqueles que combinam diferentes atributos relacionados ao perfil dos autores e/ou às características da rede e alguns propões novos atributos a serem utilizados na predição (BARTAL; SASSON; RAVID, 2009; HOSEINI; HASHEMI; HAMZEH, 2012; SUN et al., 2011; SUN et al., 2012; NARAYANAN; SHI; RUBINSTEIN, 2011).

Dentro do projeto de pesquisa descrito neste texto, tanto o problema de predição de novas coautorias quanto o problema geral de predição de coautorias foram tratados como um problema de classificação em inteligência artificial. Detalhes da abordagem utilizada são apresentados no capítulo 6.

2.13 Relação orientador-orientado

Diferentes tipos de relacionamentos podem ser considerados nas redes sociais, por exemplo, o relacionamento de amizade em redes sociais online como o Facebook e o relacionamento de seguir ou ser seguido no Twitter. Em análise de redes sociais acadêmicas, o relacionamento mais comumente usado é o de coautoria (colaboração na produção de uma publicação), porém outros relacionamentos também são estudados, como o de orientação (relação orientador-orientado), coparticipação em projetos, relação professor/aluno, colegas de trabalho, etc.

Dentro da carreira acadêmica, a orientação de alunos em seus diferentes níveis (desde trabalhos de iniciação científica ou de conclusão de curso até as orientações de alunos de doutorado) é considerada uma atividade muito importante, sendo uma das principais atividades para formação de novos pesquisadores. Adicionalmente, alguns trabalhos analisaram a relação entre produtividade do orientador e a quantidade de orientados nos diferentes tipos de orientação. Já outros, investigaram o desenvolvimento da carreira do orientado como consequência do processo de orientação (LONG; MCGINNIS, 1985; PINHEIRO; MELKERS; YOUTIE, 2014).

Nos últimos 30 anos, mais de 100 mil doutores foram titulados no Brasil[18]. A avaliação dos programas de pós-graduação realizada pela CAPES engloba tanto a avaliação geral da produção bibliográfica do corpo docente dos programas, mas também a produção bibliográfica que envolve os alunos. Dentro da pós-graduação, a atividade de pesquisa é fundamental para os alunos (e em especial aos alunos de doutorado) e uma das atividades fundamentais desse processo é a escrita e publicação dos resultados. Este tipo de habilidade desenvolvida durante a pós-graduação tem sido considerada muito importante para o bom desenvolvimento na carreira do (ex-)orientado (PINHEIRO; MELKERS; YOUTIE, 2014).

Adicionalmente, observou-se que dois aspectos da produção bibliográfica do orientado de doutorado (durante seu período de doutoramento) estão mudando ao longo do tempo: a quantidade de publicações está aumentando e o tempo entre o início do doutoramento e a primeira publicação está diminuindo (IGAMI; BRESSIANI; MUGNAINI, 2014). A crescente participação dos alunos de pós-graduação na produção científica em diferentes áreas do conhecimento foi objeto de estudo de trabalhos no Brasil (RAMOS et al., 2009; SACARDO; HAYASHI, 2011) e em diversos outros países (LEE, 2000; SALMI; GANA; MOUILLET, 2001; FRKOVIC; SKENDER; DOJCINOVIC, 2003; ANWAR, 2004; MALLETTE, 2006; ARRIOLA-QUIROZ et al., 2010; LARIVIÈRE, 2012). Muitos destes trabalhos utilizaram técnicas de resolução de entidades para identificar a relação de orientação e/ou para classificar uma produção bibliográfica como relacionada ou não ao projeto de doutorado de cada aluno. Também é comum nesse tipo de estudo utilizar uma janela temporal para investigar se uma dada produção é consequência ou não do projeto de doutorado (por exemplo, analisar as produções ocorridas cinco anos antes e cinco anos depois da data de titulação (MUGNAINI; IGAMI; BRESSIANI, 2011)).

[18] estatico.cnpq.br/painelLattes/evolucaoformacao

Em particular, há diversos trabalhos que propõe diferentes alternativas para a identificação automática da relação orientador-orientado (quando este tipo de informação não está disponível). A maioria destes trabalhos utilizam ferramentas de mineração de dados e aprendizado de máquina para inferir a existência dessa relação, pois tipicamente dispõem apenas de informações sobre produções bibliográficas (WU; CHEN; HAN, 2007; WU; CHEN; HAN, 2010; WANG et al., 2010; WANG et al., 2012).

Diferentes tipos de análises foram realizados explorando a relação orientador-orientado e a produção bibliográfica resultante dessa colaboração. Na seção 7.3 são apresentados resultados sobre a análise da participação dos orientados na produção científica de seu orientador considerando a área de Ciência da Computação (DIGIAMPIETRI; MUGNAINI; ALVES, 2013). Também foram realizadas outras análises visando a quantificar a dependência científica do orientado em relação a seu orientador ao longo do tempo (TUESTA et al., 2012; TUESTA et al., 2015a; TUESTA et al., 2015b).

2.14 Conclusões

Este capítulo apresentou os principais conceitos utilizados no desenvolvimento da pesquisa realizada e alguns dos trabalhos correlatos relacionados a cada um dos assuntos da pesquisa.

Ao longo dos próximos capítulos, serão detalhados alguns dos resultados obtidos pela análise da rede social acadêmica brasileira, tanto do ponto de vista de uma caracterização da rede quanto de novas estratégias (por exemplo, extensões de algoritmos ou combinações) utilizadas para realizar algum tipo de atividade relacionada à análise de redes sociais (por exemplo, resolução de entidades, predição de relacionamentos, e análise de tendências).

Será dado enfoque à apresentação dos resultados já publicados pelo autor em colaboração com seus orientados e colegas de trabalho, mas alguns resultados inéditos também serão apresentados.

3 Obtenção e organização dos dados

Dentro da metodologia utilizada para desenvolver cada um dos estudos, a primeira atividade consistiu da revisão bibliográfica (cujos resultados foram sumarizados no capítulo 2), seguida da obtenção, organização e enriquecimento dos dados.

Os dois tipos de arquivos obtidos da Plataforma Lattes foram utilizados (no formato XML e no formato HTML). Os currículos da Plataforma Lattes podem ser obtidos pela internet, utilizando-se, por exemplo, o comando *wget* ou a ferramenta *scriptLattes*. Porém, para isso, é necessário encontrar o identificador único de cada currículo, utilizado para compor a URL (*Uniform Resource Locator*) completa do currículo.

A primeira seção deste capítulo apresenta as estratégias utilizadas para a obtenção dos identificadores dos currículos. Após esta identificação, os currículos podem ser copiados e as informações de interesse extraídas (seção 3.2). Na seção 3.3 é apresentado o banco de dados relacional que foi construído para armazenar as informações consideradas pertinentes dos Currículos Lattes. A seção 3.4 descreve outras informações (disponíveis na internet ou mesmo implícitas nos currículos) que podem ser úteis para complementar análises bibliométricas ou de redes sociais acadêmicas e que foram utilizadas nas pesquisas apresentadas neste trabalho. Este capítulo é encerrado com a apresentação de resultados sobre a análise da atualização dos Currículos Lattes (seção 3.5).

3.1 Obtenção dos identificadores dos currículos

Neste projeto, para encontrar os identificadores dos currículos, três estratégias foram utilizadas em diferentes momentos e com diferentes finalidades.

A **primeira estratégia** estava contextualizada num projeto mais amplo de obtenção de dados via Web e partiu do princípio da busca de dados na internet visível por meio do uso de motores de busca (DIGIAMPIETRI; SILVA, 2011). Foram utilizados os motores de busca da Google[1] e da Microsoft (Bing[2]), que receberam como *string* de busca o nome completo do pesquisador mais a expressão "currículo Lattes". Os resultados retornados pelos motores de busca são então minerados procurando-se pela URL de currículos Lattes. Para toda URL encontrada, verifica-se se ela já está na base de currículos criada para o

[1] https://www.google.com.br
[2] http://www.bing.com/

projeto e, caso contrário, o respectivo currículo é copiado e é feita uma validação para verificar se o currículo corresponde ao nome que foi procurado. Independente do resultado da verificação, a URL do currículo e o nome de seu possuidor são guardados na base de currículos.

Esta estratégia apresenta duas limitações principais: exige uma validação manual para evitar homônimos e é necessário que os currículos sejam encontrados pelos motores de busca utilizados. Uma validação do sistema envolvendo o nome completo de 1.002 professores de diferentes universidades (DIGIAMPIETRI; SILVA, 2011) apresentou os seguintes resultados. O sistema encontrou 974 URLs de currículos Lattes, destas 822 foram identificadas corretamente como URLs dos currículos procurados (82% dos procurados); 136 foram identificadas corretamente como não pertencentes aos docentes procurados; não houve nenhum caso falso-positivo e ocorreram 16 casos de falso-negativos (o sistema encontrou o currículo procurado, mas não reconheceu que ele pertencia ao respectivo professor [devido a variações entre o nome cadastrado no currículo e o nome completo utilizado na busca]).

Para cada currículo encontrado, é possível extrair os identificadores de alguns de seus vizinhos (coautores, orientadores ou orientandos) utilizando-se os *links* explícitos presentes em cada currículo. Uma limitação desta abordagem é que não há *links* explícitos para todos os coautores, orientadores e orientandos que possuem currículos Lattes. Uma maneira mais completa de se identificar os vizinhos é analisar toda a base de currículos Lattes e utilizar algumas estratégias de resolução de entidades (conforme será apresentado no capítulo 4.

É possível encontrar currículos Lattes a partir do nome do aluno/pesquisador a partir da interface de busca da Plataforma Lattes. Esta API foi utilizada na segunda estratégia, porém com o objetivo de encontrar um grande número de currículos e não apenas de uma lista específica de pessoas.

A **segunda estratégia** para encontrar os identificadores dos currículos foi dividida em duas etapas (DIGIAMPIETRI et al., 2012a). Na primeira, foram feitas 80 consultas à interface provida pela Plataforma Lattes, mas ao invés de buscar por nomes de pesquisadores, cada consulta utilizava o nome de cada uma das (sub-)áreas do conhecimento do CNPq como palavras-chave da busca por assunto. Por exemplo, a busca por “Ciência da

Computação" retorna 2.600 currículos[3]. Uma ferramenta foi desenvolvida para extrair os identificadores do resultado destas consultas. Mais de 100.000 currículos foram identificados nesta etapa. Cada um destes currículos foi copiado e, na segunda etapa, foram procurados identificadores de currículos dentro de cada um destes currículos (o que está sendo chamado neste trabalho de *links* explícitos a outros currículos). Esta segunda etapa foi executada de maneira iterativa (sempre que novos currículos eram encontrados, verificava-se se eles possuíam identificadores de currículos ainda não encontrados).

A combinação destas duas estratégias possibilitou, em maio de 2011, a identificação de 1.236.548 currículos que foram copiados em seu formato HTML, totalizando pouco mais de 16 GB de dados. Estas estratégias não objetivaram a obtenção de todos os currículos da Plataforma Lattes, mas sim de um conjunto significativo de currículos para servirem de base para a criação e análise de redes sociais acadêmicas e, em especial, todos os currículos pertencentes ao componente gigante da rede (considerando-se os *links* explícitos) (DIGIAMPIETRI et al., 2012a). Apenas para adicionar informações sobre a quantidade de currículo, em 2007 o CNPq anunciou que a base da Plataforma Lattes atingiu um milhão de currículos.

A **terceira estratégia** surgiu após uma verificação incidental dos resultados de busca utilizando-se a interface de busca da Plataforma Lattes. Observou-se que ao fazer uma busca passando como parâmetro um espaço em branco (e ativando os parâmetros demais pesquisadores [além dos doutores] e realizando a busca por assunto) a interface retorna todos os currículos que possuem ao menos um espaço em branco em seu conteúdo. Esta consulta retornou, por exemplo, no dia 14 de julho de 2015, 4.262.493 resultados (4,2 milhões de currículos). Esta estratégia foi utilizada para a obtenção do que consideramos todos os currículos da Plataforma Lattes (este processo foi realizado duas vezes: no primeiro semestre de 2013, obtendo-se pouco menos de 3,2 milhões de currículos e no primeiro semestre de 2015, obtendo-se pouco menos de 4,2 milhões de currículos).

3.2 Processamento inicial dos currículos

Os arquivos HTML e XML dos currículos foram processados de maneiras diferentes. Para processar os arquivos HTML, foi desenvolvido um *parser* que utiliza as marcações

[3] Consulta realizada em 14 de julho de 2015 usando a expressão "Ciência da Computação" entre parênteses e buscando por currículos de doutores e não doutores.

HTML (*tags*) para separar cada currículo em seções, subseções, itens e assim por diante. Infelizmente, este tipo de processamento não é totalmente preciso devido à existência de diversos campos opcionais e ao fato de não haver um delimitador específico para alguns campos. Assim, é impossível desenvolver uma ferramenta que identifique corretamente todos os campos (por exemplo, uma publicação em anais pode ter em seu título um ou mais pontos e pontos são a única delimitação entre o título da publicação e o nome do evento, bem como entre a descrição do evento e o título dos anais (sendo que estes campos também podem ter pontos internos). Apesar deste tipo de limitação, a maioria dos campos pode ser corretamente identificada. Após a identificação das seções, subseções, itens e campos de cada currículo é gerado um arquivo XML com os itens considerados relevantes para as análises desenvolvidas neste trabalho.

Os arquivos XML resultantes do processamento dos arquivos HTML são então utilizados como entrada para um programa que cria um banco de dados relacional com os dados dos currículos.

3.3 Banco de dados relacional

Como parte do processo de organização dos dados dos currículos foi criado um banco de dados relacional. O sistema gerenciador de banco de dados (SGBD) utilizado foi o PostgreSQL[4].

A tabela central do banco de dados é a tabela *Currículos* que contém informações gerais como identificador, nome, tipo de bolsa produtividade, gênero e data da última atualização do currículo. A tabela *Publicações* armazena os dados gerais dos diferentes tipos de publicações cadastradas na Plataforma Lattes. Há sete tipos de publicações que foram considerados: artigos completos publicados em periódicos; artigos aceitos para publicação; trabalhos completos publicados em anais de congressos; resumos expandidos publicados em anais de congressos; resumos publicados em anais de congressos; livros publicados organizados ou edições; e capítulos de livros publicados.

A tabela *Formações* contém as informações sobre formação e titulação, incluindo o período de formação, o título, a instituição, o identificador do orientador e o nome do orientador. As áreas de atuação são armazenadas na tabela *AreasDeAtuação* da mesma maneira que são estruturadas no currículo: grande área, área, subárea e especialidade.

[4] www.postgresql.org/

Quatro tabelas contêm as informações relacionadas a projetos de pesquisa: *ProjetosDePesquisa*, *ProjetosDePesquisa_descricao*, *CoordenadoresProjetos* e *IntegrantesProjetos*. A tabela *Orientações* contém a lista dos diferentes tipos de orientação realizados pelo pesquisador. Sete tipos de orientação foram considerados: Orientações de Pós-Doutorado; Teses de Doutorado; Orientações de Outra Natureza; Dissertações de Mestrado; Monografias de Conclusão de Curso de Aperfeiçoamento; Iniciações Científicas; e Trabalhos de Conclusão de Graduação. O diagrama entidade relacionamento, bem como detalhes adicionais destas tabelas podem ser encontrados em Digiampietri et al. (2012a).

Os 1.236.548 currículos encontrados de acordo com a segunda estratégia de identificação das URLs dos currículos (ver subseção 3.1) foram convertidos para arquivos XML e então utilizados para popular o banco de dados relacional. O banco de dados resultante possui (sem a eliminação de eventuais registros duplicados): 1.378.885 projetos de pesquisa; 3.250.846 registros sobre formação/titulação; 3.256.019 registros de áreas de atuação (isto é, na média, cada currículo contém a declaração de 2,6 áreas/subáreas de atuação), 4.329.993 registros de orientações; e 11.529.218 publicações. Destaca-se aqui que estes 11,5 milhões de registros de publicações contêm redundâncias que foram posteriormente tratadas utilizando-se técnicas de resolução de entidades (seção 4).

Algumas tabelas adicionais foram criadas para permitir ou facilitar algumas análises. Por exemplo, foram armazenados alguns índices ou métricas relacionados a cada publicação (número de citações recebidas, Qualis do veículo no qual o artigo foi publicado, etc) e também sobre cada autor (como índice g e índice h). Além disto, para parte do conjunto de dados foram inseridas as informações sobre o credenciamento do pesquisador em programas de pós-graduação (indicando qual o programa e qual o período do credenciamento, quando disponível). Informações sobre os dados utilizados para enriquecer a base de Currículos Lattes serão apresentadas na próxima seção.

3.4 Enriquecimento do conjunto de dados

Diferentes tipos de informação foram utilizados para enriquecer os dados dos currículos Lattes: dados da pós-graduação; dados de veículos de publicação; citações; e métricas derivadas. Não foi realizado o enriquecimento de todo o conjunto de dados, pois nem todas as informações apresentadas a seguir foram consideradas necessárias para todas as análises realizadas.

Os dados referentes aos *programas de pós-graduação* foram obtidos dos relatórios da avaliação realizada pela CAPES foram utilizados para identificar quais docentes estavam credenciados em cada programa de pós-graduação, qual a nota do programa e o período de credenciamento. Estes relatórios contêm apenas o nome completo do docente e por isso uma estratégia de resolução de nomes foi necessária para identificar o currículo correspondente (este tipo de estratégia será discutido no capítulo 4). Na presente pesquisa só foram utilizados os dados dos programas de pós-graduação em Ciência da Computação avaliados nos triênios 2004-2006 e 2007-2009.

Três tipos de índices relacionados aos *veículos de publicação* foram utilizados: fator de impacto JCR (*Thompson's Journal Citation Reports*), índice SJR (*Scimago Journal Rank*) e Qualis. Para cada um destes, foram copiadas as tabelas da internet contendo toda a classificação e convertidas para arquivos no formato CSV. O cruzamento entre a informação oriunda dos Currículos Lattes e as informações destes índices ocorreu por meio de casamento direto (exato) no caso dos artigos em periódicos com ISSN e aproximado para o casamento do nome dos eventos. Diferentes estratégias para resolução de nomes e comparação aproximada de nomes/*strings* serão apresentadas no próximo capítulo.

Os índices dos veículos de publicação foram associados às publicações para a realização de algumas análises bibliométricas. Este tipo de prática é comum neste tipo de análise e visa a qualificar de maneira automática as publicações atribuindo a elas o índice do veículo no qual elas foram publicadas. Vale destacar que as medidas não foram feitas para qualificar artigos, pesquisadores ou grupos de pesquisa.

Para artigos completos publicados (em revistas ou anais de eventos), duas informações foram obtidas: o total de citações extraído do *site* do Google Scholar[5] e o total de citações extraído do *site* Microsoft Academic Search[6].

Por fim, algumas medidas derivadas foram calculadas. Por exemplo, é possível criar uma medida chamada citações associada a uma pessoa (ou a um grupo de pessoas) contendo o total de citações recebidas pelas publicações da pessoa (ou do grupo). Isto pode ser feito para qualquer uma das medidas, e o número total de publicações, por si só, pode ser uma dessas medidas. Adicionalmente, a partir das informações sobre as citações é possível calcular os índices g e h (tanto para uma única pessoa quanto para um grupo de pessoas).

[5] scholar.google.com.br
[6] http://academic.research.microsoft.com/

3.5 Análise da atualização dos dados

Conforme apresentado, três fatores são considerados extremamente relevantes para a análise de dados: corretude, completude e atualização (CANIBANO; BOZEMAN, 2009). É possível, de forma automática ou semiautomática realizar alguns estudos para analisar cada uma destas questões dentro dos currículos da Plataforma Lattes.

Por exemplo, é possível extrair uma medida de corretude comparando-se registros de coautores e verificando se há diferenças importantes nas informações cadastradas ou comparando-se registros de alguma base curada (por exemplo, uma base de dados institucional) com os dados dos currículos.

Um procedimento semelhante pode ser feito para verificar a completude e a atualização (desde que se tenha algum conjunto de dados de referência atualizado).

Dentro deste contexto, na presente pesquisa foi analisada a data da última atualização dos Currículos Lattes. Foi utilizado um conjunto de mais de três milhões de currículos e estes foram agrupados de acordo com a formação acadêmica e a grande área de atuação. Além da simples observação do período transcorrido desde a última atualização de cada currículo e data em que os currículos foram copiados, neste estudo se estimou a quantidade de informação que poderia estar ausente em termos de quantidade de publicações com base na produção média de cada pessoa nos últimos anos e no período transcorrido desde a última atualização.

Observou-se que, de um modo geral, quanto maior a formação/titulação das pessoas mais recentemente os currículos foram atualizados pela última vez. A estimativa sobre a completude dos registros das publicações indicou a possibilidade de que mais de 20% dos registros de artigos publicados nos 36 meses anteriores a data em que os currículos foram copiados possam estar faltando (DIGIAMPIETRI et al., 2014a; DIGIAMPIETRI et al., 2014b).

4 Resolução de nomes

Conforme apresentado na seção 2.6, a resolução de nomes é uma atividade importante na análise de redes sociais acadêmicas. Dentro do presente projeto de pesquisa, diferentes estratégias foram utilizadas para a resolução do nome dependendo dos tipos e da quantidade de dados disponíveis. Assim como na maioria dos trabalhos que tratam de dados preenchidos manualmente (como é o caso dos dados dos currículos da Plataforma Lattes), as estratégias desenvolvidas utilizaram alguns tipos de aproximação (OKAZAKI; TSUJII, 2010; COHEN; RAVIKUMAR; FIENBERG, 2003) para a resolução de nomes. Estas estratégias são apresentas nas próximas seções.

4.1 Resolução de publicações

A informação mais simples e tipicamente mais disponível de referências a publicações são os títulos. Quando apenas este tipo de informação está disponível, algumas das formas de se comparar essas referências para verificar se estão relacionadas à mesma entidade são: comparação direta dos caracteres dos títulos, verificação se um dos títulos é uma *substring* do outro, comparação aproximada dos títulos (usando algum tipo de medida de distância e um limiar), entre outras. Além disso, há alguns pré-processamentos comumente executados sobre os títulos: colocar todas as letras em maiúsculas (ou minúsculas), remoção de pontuações, remoção de acentos, remoção de *stop-words*, entre outros.

A tabela 5 ilustra os resultados de algumas estratégias simples para a resolução de publicações considerando-se apenas o título. O experimento objetivava verificar a quantidade de arestas que cada estratégia consegue recuperar considerando que, cada pesquisador é um nó de uma rede e uma publicação corresponde a um clique envolvendo todos os coautores desta publicação na rede. O conjunto de dados utilizados foi extraído dos Currículos Lattes dos professores permanentes do programa de pós-graduação em Ciência da Computação do Instituto de Matemática e Estatística da Universidade de São Paulo[1] considerando o triênio 2007-2009 e corresponde aos artigos publicados em periódicos. Este conjunto de dados é composto por 486 artigos diferentes, sendo que 96 são em coautoria com outro professor do programa. Ao todo, estas publicações produziriam 205 arestas no grafo de colaborações (permitindo-se arestas múltiplas quando um par de

[1] http://www.ime.usp.br/

pesquisadores colaborou na publicação de mais de um artigo). O experimento realizado visa a verificar quantas destas arestas foram corretamente identificadas e se ocorreu a identificação incorreta de alguma aresta. Foram analisadas as seguintes métricas: VP (Verdadeiro-Positivo) - número de arestas identificadas; FP (Falso-Positivo) número de arestas inexistentes que foram incorretamente criadas; Precisão - porcentagem das arestas corretamente encontradas em relação ao número total de arestas identificadas; Revocação - porcentagem de arestas corretamente encontradas em relação ao número real de arestas.

Os títulos utilizados estavam todos em letras minúsculas e foram testadas duas condições: sem realização de nenhum pré-processamento adicional ou executando um filtro que removia os acentos e pontuações. Além disso, três estratégias foram utilizadas: comparação exata dos títulos; verificação se um título é uma *substring* do outro (esta verificação auxilia a recuperar os casos em que um autor cadastrou o título e o subtítulo do artigo enquanto o outro cadastrou apenas o título); comparação aproximada de títulos utilizando a distância de edição proposta por Levenshtein (1966) (neste caso foram usados diferentes limiares tanto em valores absolutos quanto em porcentagem em relação ao tamanho do menor título entre os dois que estão sendo comparados).

As células das colunas correspondentes às medidas de Precisão e Revocação da tabela 5 estão coloridas de acordo com seu valor (maiores valores estão em tons mais intensos de verde enquanto menores valores estão em tons mais intensos de vermelho). Conforme era de se esperar, na estratégia que usa distância de edição, ao se aumentar o valor do limiar da distância para que duas referências a títulos sejam consideradas uma mesma entidade, ocorre um aumento na revocação com eventual diminuição da precisão.

Ao se fixar a precisão em 100%, é possível observar nesse pequeno experimento que o melhor valor para a revocação foi de 88,8%, alcançado ao se utilizar títulos filtrados e com distância de edição máxima igual a 6. Este valor foi muito próximo aos valores obtidos utilizando-se também os títulos filtrados, mas com distância de edição máxima igual a 5 ou com distância de edição relativa igual a 10% ou 12,5% do título de menor tamanho (entre os títulos filtrados que estão sendo comparados).

Diversos trabalhos utilizam um limiar próximo a 10% para a resolução de publicações utilizando seus títulos. Por exemplo, na ferramenta *scriptLattes* (MENA-CHALCO; CESAR-JUNIOR, 2009) o valor utilizado é de 8%, porém é importante destacar que nessa ferramenta a distância de edição não é o único critério utilizado.

Tabela 5 – Resultados para a resolução de publicações considerando apenas o título

Estratégia	Pré-processamento	VP	FP	Precisão	Revocação
Casamento exato de títulos	nenhum	164	0	100,0%	80,0%
Casamento exato de títulos	filtro de acentos e pontuações	167	0	100,0%	81,5%
Substring dos títulos	filtro de acentos e pontuações	181	11	94,3%	88,3%
Distância de edição 1	nenhum	169	0	100,0%	82,4%
Distância de edição 2	nenhum	171	0	100,0%	83,4%
Distância de edição 3	nenhum	174	0	100,0%	84,9%
Distância de edição 4	nenhum	177	0	100,0%	86,3%
Distância de edição 5	nenhum	178	0	100,0%	86,8%
Distância de edição 6	nenhum	179	0	100,0%	87,3%
Distância de edição 7	nenhum	180	1	99,4%	87,8%
Distância de edição 8	nenhum	181	5	97,3%	88,3%
Distância de edição 9	nenhum	183	8	95,8%	89,3%
Distância de edição 10	nenhum	183	15	92,4%	89,3%
Distância de edição 1	filtro de acentos e pontuações	174	0	100,0%	84,9%
Distância de edição 2	filtro de acentos e pontuações	177	0	100,0%	86,3%
Distância de edição 3	filtro de acentos e pontuações	178	0	100,0%	86,8%
Distância de edição 4	filtro de acentos e pontuações	180	0	100,0%	87,8%
Distância de edição 5	filtro de acentos e pontuações	181	0	100,0%	88,3%
Distância de edição 6	filtro de acentos e pontuações	182	0	100,0%	88,8%
Distância de edição 7	filtro de acentos e pontuações	183	1	99,5%	89,3%
Distância de edição 8	filtro de acentos e pontuações	183	5	97,3%	89,3%
Distância de edição 9	filtro de acentos e pontuações	183	9	95,3%	89,3%
Distância de edição 10	filtro de acentos e pontuações	183	16	92,0%	89,3%
Distância de edição 2,5%	nenhum	173	0	100,0%	84,4%
Distância de edição 5%	nenhum	176	0	100,0%	85,9%
Distância de edição 7,5%	nenhum	178	0	100,0%	86,8%
Distância de edição 10%	nenhum	181	0	100,0%	88,3%
Distância de edição 12,5%	nenhum	181	0	100,0%	88,3%
Distância de edição 15%	nenhum	182	2	98,9%	88,8%
Distância de edição 17,5%	nenhum	183	10	94,8%	89,3%
Distância de edição 20%	nenhum	184	11	94,4%	89,8%
Distância de edição 22,5%	nenhum	186	11	94,4%	90,7%
Distância de edição 25%	nenhum	189	11	94,5%	92,2%
Distância de edição 2,5%	filtro de acentos e pontuações	177	0	100,0%	86,3%
Distância de edição 5%	filtro de acentos e pontuações	179	0	100,0%	87,3%
Distância de edição 7,5%	filtro de acentos e pontuações	180	0	100,0%	87,8%
Distância de edição 10%	filtro de acentos e pontuações	181	0	100,0%	88,3%
Distância de edição 12,5%	filtro de acentos e pontuações	181	0	100,0%	88,3%
Distância de edição 15%	filtro de acentos e pontuações	183	2	98,9%	89,3%
Distância de edição 17,5%	filtro de acentos e pontuações	184	11	94,4%	89,8%
Distância de edição 20%	filtro de acentos e pontuações	186	11	94,4%	90,7%
Distância de edição 22,5%	filtro de acentos e pontuações	190	11	94,5%	92,7%
Distância de edição 25%	filtro de acentos e pontuações	190	11	94,5%	92,7%

Fonte: Digiampietri (2015)

Uma estratégia um pouco mais sofisticada desenvolvida no presente projeto de pesquisa consiste da análise de três condições (DIGIAMPIETRI et al., 2012b). Duas referências a artigos extraídas da Plataforma Lattes são consideradas referências a um mesmo artigo (à mesma entidade) se três condições forem satisfeitas: (i) os títulos forem compatíveis; (ii) a lista de autores for compatível; e (iii) as demais informações forem compatíveis. Cada uma dessas condições é descrita na tabela 6.

Tabela 6 – Critérios utilizados para a resolução de títulos

Condição	Descrição
Condição 1	Dois títulos de produções bibliográficas são considerados compatíveis se são iguais; *OU* se a diferença entre o tamanho dos dois títulos for menor do que um terço da soma do tamanho dos títulos *E* {ambos possuem mais de 10 caracteres e um estiver contido dentro do outro *OU* a Distância Levenshtein (LEVENSHTEIN, 1966) entre os dois títulos for menos do que 5}. Obviamente a última parte da condição garantiria a primeira parte, porém, devido à maior complexidade computacional necessária para calculá-la, a verificação de compatibilidade de título é executada na ordem apresentada.
Condição 2	Duas lista de (co)autores são consideradas compatíveis se, ao se comparar as duas listas, houver mais autores em comum do que diferentes, considerando-se apenas o casamento exato do último sobrenome de cada autor. Foi considerado o último sobrenome por, na maioria dos casos, ser invariante a abreviações.
Condição 3	As demais informações serão compatíveis se ao menos dois dos seguintes quatro campos forem iguais: ano de publicação, local, páginas e volume.

Fonte: adaptado de Digiampietri et al. (2012b)

Os valores utilizados nas condições da tabela 6 foram definidos após testes utilizando um conjunto de treinamento contendo 330 referências a publicações (DIGIAMPIETRI et al., 2012b) englobando os sete tipos presentes na Plataforma Lattes: artigos completos publicados em periódicos; artigos aceitos para publicação; trabalhos completos publicados em anais de congressos; resumos expandidos publicados em anais de congressos; resumos publicados em anais de congressos; livros publicados organizados ou edições; e capítulos de livros publicados.

A estratégia foi validada utilizando-se o conjunto de dados contendo 486 artigos diferentes (apresentados no início desta seção). Como resultados, foram identificados corretamente 468 (taxa de verdadeiro-positivos igual a 96,3% do total de artigos); 5 artigos foram identificados como únicos quando na verdade eram artigos diferentes (falso positivos) e a estratégia deixou de unir 36 registros que correspondiam a 18 artigos diferentes (falso negativos).

O cálculo da distância de edição (LEVENSHTEIN, 1966) utilizando programação dinâmica tem complexidade computacional da ordem de $m * n$, sendo m e n os tamanhos das sequências de caracteres que estão sendo comparadas. Caso se deseje comparar cada um dos títulos de um conjunto de referências a artigos contra os demais e assumindo-se que há t títulos haverá $t * (t-1)$ comparações. Supondo-se que cada título possui 20 caracteres, a complexidade de se comparar t títulos será dada por $400 * t * (t-1)$. Caso se deseje comparar 10 milhões de títulos da Plataforma Lattes, o número de operações realizadas será da ordem de $4 * 10^{16}$. Imaginando-se que é possível realizar 1 bilhão de operações

por segundo, será necessário $4 * 10^7$ segundos (666.667 minutos, ou 11.111 horas, ou 463 dias), o que potencialmente tornaria esse tipo de cálculo impraticável. Existem diferentes maneiras de reduzir o tempo total gasto, por exemplo, as comparações entre títulos são atividades independentes então poderiam ser realizadas em paralelo. Assim, em um *grid* com mil núcleos é possível resolver este problema em menos de 12 horas (assumindo-se que cada núcleo possa processar 1 bilhão de operações por segundo).

Outra alternativa utilizada corresponde a não comparação de todos os pares de títulos, que pode ser baseada simplesmente no próprio título ou em informações adicionais. Por exemplo, na análise de mais de um milhão de Currículos Lattes (MENA-CHALCO; CESAR-JUNIOR, 2009) um dos critérios foi só comparar títulos que iniciam com a mesma letra (após o pré-processamento dos títulos que pode ou não envolver a remoção de *stop-words*). Este tipo de critério diminui o tempo total de processamento pelo fato de que o número de comparações entre títulos é proporcional ao quadrado do número de título. Assim, ao se agrupar os títulos pelas letras iniciais, teremos 27 grupos menores (considerando as 26 letras e um grupo adicional para os títulos que não iniciam por uma letra). No melhor caso, os 27 grupos teriam a mesma quantidade de títulos e o processamento levaria cerca de 1/27 do tempo total sem o agrupamento (pois ocorreriam 27 processamentos [um para cada grupo], cada um levando cerca de 1/(27*27) do tempo do processamento total). No pior caso não haveria ganho de tempo, pois todos os títulos seriam iniciados pela mesma letra (MENA-CHALCO; CESAR-JUNIOR, 2009). Nos testes que avaliaram a distribuição dos títulos de acordo com a letra inicial, foi observado que o tempo de processamento seria reduzido para cerca de 14% do tempo total sem esta divisão em grupos. Além disso, ao se dispor de informações adicionais, outros critérios são utilizados: só comparar referências de artigos que possuem o mesmo ano de publicação (ou no máximo a diferença de um ano, considerando possíveis cadastros imprecisos); comparar apenas referências de artigos do mesmo tipo (isto pode causar alguns problemas pelo fato de autores cadastrarem de diferentes formas o mesmo artigo); etc. Há ainda uma grande gama de opções para o agrupamento de títulos antes de se calcular a distância de edição, utilizando, por exemplo, estratégias de indexação.

4.2 Resolução de nomes de autores

Dentro do presente projeto de pesquisa, duas estratégias foram utilizadas relacionadas à identificação ou à desambiguação do nome de autores. A primeira tem por objetivo, dado o nome completo de uma pessoa que potencialmente seja autora de uma publicação, encontrar o nome citado desta pessoa entre a lista de autores de uma publicação. A segunda estratégia tem por objetivo, dada uma lista de referências a publicações, identificar quais são os autores considerando-se as diferentes formas em que são citados.

4.2.1 Primeira estratégia

A primeira estratégia é utilizada dentro de um contexto de referências a artigos mais fechado no qual se sabe que uma dada pessoa é (ou deveria ser) autora de um dado artigo ou ao menos têm uma chance relativamente grande de ser. Por exemplo, nos registros do Banco de Dados Bibliográficos da Universidade de São Paulo (Dedalus[2]), para cada registro de publicação envolvendo docentes da universidade há o cadastro do nome completo dos docentes (extraído do Sistema de apoio à avaliação e a gestão institucional da USP [Tycho[3]]) e o cadastro da lista de autores extraída da publicação propriamente dita. Até 2009 o sistema Dedalus continha um banco de autoridades de autor para registro normalizado dos nomes de autores USP, porém, com a migração dos dados em 2010 este banco foi descontinuado (MUGNAINI et al., 2012).

O conjunto de dados contém uma lista contendo um ou mais nomes completos de docentes e uma lista de autores da maneira que seus nomes são citados (por exemplo, sobrenome e iniciais dos demais nomes). Exceto por algum problema de cadastramento, é esperado que a primeira lista esteja contida na segunda. O objetivo da estratégia desenvolvida para a identificação dos nomes dos autores foi, dado o nome completo do autor, encontrar seu nome citado na publicação de forma a identificar em que posição ele foi citado (primeiro autor, segundo, etc) e verificar possíveis problemas de cadastramento. A mesma estratégia (com mudança em apenas alguns parâmetros) foi empregada para verificar a participação dos orientados na produção de seus orientadores (utilizando dados da Plataforma Lattes).

[2] http://dedalus.usp.br/
[3] https://uspdigital.usp.br/tycho

A estratégia desenvolvida utiliza quatro critérios, sendo que cada critério só é avaliado caso o anterior tenha falhado. Os critérios variam da busca exata pelo nome completo do autor até busca aproximada pelo sobrenome e alguns dos outros nomes (inicial e nomes do meio), conforme descrito na tabela 7. Os nomes utilizados foram pré-processados colocando-se todas as letras em caixa alta e removendo-se os acentos.

Tabela 7 – Critérios utilizados para a identificação do autor

Critério	**Descrição**
(i)	Busca pelo nome completo do docente dentro dos registros do Dedalus exatamente da maneira que ele aparece no sistema Tycho;
(ii)	Busca pelo nome do docente, permitindo-se que um ou mais nomes do meio estejam abreviados. Exige-se assim, que ao menos o primeiro e o último nome estejam completos e que todos os demais nomes estejam completos ou abreviados (mas nenhum nome ou abreviação pode estar faltando);
(iii)	Consideraram-se os mesmos critérios anteriores, porém permitindo-se a ausência ou excesso do último sobrenome (e neste caso, exige-se que o sobrenome anterior seja encontrado). Esta estratégia foi desenvolvida para tratar principalmente dois casos: o caso das pessoas que adotam um novo sobrenome após o casamento e nomes que são encerrados por "Filho", "Júnior" e outros do gênero.
(iv)	Permite uma combinação de várias diferenças entre os nomes que estão sendo comparados. Especificamente, permite-se que o primeiro nome esteja abreviado, que haja um nome ou uma abreviação sobrando/faltando em um dos nomes buscados, que nomes não abreviados sejam considerados compatíveis caso as diferenças entre eles sejam de no máximo duas letras (utilizando-se um algoritmo de distância de edição para calcular estas diferenças). Devido a grande combinação de situações possíveis na comparação entre dois nomes, a estratégia adotada utilizou um esquema de pontuações positivas e negativas para cada situação. Por exemplo, se os dois nomes possuíssem o último sobrenome em comum, seria atribuída uma nota positiva. Por outro lado, se houvesse uma pequena diferença entre os sobrenomes (até 2 letras de diferença) seria atribuída uma nota positiva, porém menor que a primeira. O mesmo princípio de pontuação é utilizado para as diferentes situações: nomes ausentes ou em excesso, abreviações, etc. Se a pontuação final, após a comparação de todas as partes dos dois nomes completos em verificação for positiva, então o sistema considera que conseguiu encontrar o nome buscado, caso contrário considera que o nome não havia sido encontrado. Os valores das pontuações utilizados para cada situação foram estabelecidos de maneira empírica.

Fonte: adaptado de Mugnaini et al. (2012)

A estratégia foi testada considerando 12.628 registros bibliográficos, correspondendo a produção bibliográfica de 2006 a 2010 de quatro unidades da USP: Escola de Artes, Ciências e Humanidades (EACH), Escola de Comunicações e Artes (ECA), Faculdade de Educação (FE) e Instituto de Física de São Carlos (IFSC). Estas produções contêm 1.137 autores docentes diferentes. Destes, 74,2% foram identificados pelo primeiro critério (casamento exato dos nomes). Ao se considerar o número total de registros de autores docentes, há 28.284 registros (ou seja, cada um dos 1.137 autores docentes possui, em média, 24,9 produções cadastradas). Dos 28.284 registros, 92,3% foram identificados corretamente pela busca exata dos nomes (MUGNAINI et al., 2012).

Como resultados, foram identificados mais de de 99,5% dos autores e a validação manual realizada sobre 500 registros indicou uma precisão de 100% nas identificações. Dentre os casos em que a estratégia desenvolvida não foi capaz de identificar os autores, destaca-se o caso de cadastro incorreto do número de autores (em 0,1% dos registros há mais autores docentes cadastrados do que o número de nomes na lista de autores) e casos em que o nome encontrado na lista de autores não corresponde ao nome do autor, propriamente dito, e sim a um apelido ou pseudônimo (por exemplo, um registro contém o nome “Toninho” ao invés de “Antônio Augusto”).

As identificações de nomes que não utilizaram o casamento exato entre o nome completo do autor e o nome utilizado no registro da publicação foram classificadas em oito categorias de acordo com o tipo de diferença encontrada entre o nome completo e o nome nas citações: nomes a menos nas citações (nomes do meio ausentes nas citações), sobrenomes a menos nas citações, abreviações (existência de abreviações, compatíveis com o primeiro nome e/ou nomes do meio), nomes com diferenças (com palavras diferentes), nomes parecidos (erro de digitação), sobrenomes parecidos (erro de digitação), sobrenomes a mais nas citações, e nomes invertidos (ordem invertida) (MUGNAINI et al., 2012). A tabela 8 contém a contagem das ocorrências em cada uma dessas categorias.

Tabela 8 – Variações entre nome completo e nome nos registros bibliográficos

Tipos de variações de nomes encontrados	**Total de docentes**	**Total de ocorrências**
Nomes a menos	185	1.466
Sobrenomes a menos	41	376
Abreviações	79	179
Nomes com diferenças	15	56
Sobrenomes parecidos	9	43
Sobrenomes a mais	7	37
Nomes parecidos	13	35
Nomes invertidos	0	0

Fonte: Mugnaini et al. (2012)

A estratégia apresentada foi também utilizada para verificar a participação dos orientados na produção bibliográfica do orientador utilizando dados extraídos dos Currículos Lattes, porém foram utilizados diferentes pesos para o quarto critério de identificação. Os resultados deste estudo são sumarizados na seção 7.3.

4.2.2 Segunda estratégia

A segunda estratégia desenvolvida visa, partindo-se das referências das publicações, a identificar todas as obras de cada um dos autores. Este problema é o mais comumente tratado na desambiguação de nomes de autores.

Para esta estratégia, foram utilizados dados do projeto DBLP[4] (*Digital Bibliography & Library Project*), que consiste de uma fonte de referências online para produções acadêmicas na área de Ciência da Computação. Os dados podem ser acessados diretamente pela página do projeto ou é possível copiar toda a base do projeto em um arquivo compactado contendo arquivos XML com informações sobre as publicações como: lista de autores, título do artigo, ano de publicação, número de páginas, nome do periódico ou do evento.

Atualmente, o DBLP conta com mais de 3 milhões de registros de publicações[5], sendo a maioria referente a artigos publicados em anais de eventos. Sempre que uma nova lista de artigos é inserida no projeto (correspondendo, por exemplo, aos anais de um evento ou as publicações de um número de um periódico) há uma verificação se os autores destas novas publicações já existem no DBLP, porém este processo ainda pode ser melhorado[6].

Atualmente os dados do DBLP contam com 1.594.110 registros de autores (já considerando o processo de resolução de nomes empregado pelos responsáveis). Entretanto, ainda há diferentes registros que se referem a um mesmo autor.

Dentro do presente projeto de pesquisa, objetivou-se desenvolver uma estratégia de desambiguação de nomes de autores que aperfeiçoe este processo considerando os dados do DBLP.

Conforme apresentado na seção 2.6, há diferentes estratégias para a desambiguação de nomes, entre elas: cálculos de distância ou similaridades entre *strings*, por exemplo distância de edição e distância cosseno; uso de mineração de texto para tentar atribuir áreas/assuntos aos autores; uso de métricas de redes sociais (como vizinhos em comum); uso de técnicas de *agrupamento* ou *blocking* para agrupar registros de autores semelhantes (estratégia não supervisionada); uso de funções de regressão para identificar se um par de

[4] http://dblp.uni-trier.de/

[5] Informação obtida em 30/07/2015, no site do projeto.

[6] No dia 30/07/2015 era possível encontrar no site do DBLP: http://dblp.uni-trier.de/db/ o anúncio de uma vaga para contratar um pesquisador para trabalhar no projeto *Scalable Author Disambiguation for Bibliographic Databases* que visa a melhorar o processo de desambiguação de nomes de autores utilizado no projeto DBLP.

referências de autores se refere ao mesmo autor (utilizando-se um limiar de distância ou similaridade); e uso de técnicas de classificação para identificar se um par de referências de autores corresponde a um mesmo autor (estratégia supervisionada) (DIGIAMPIETRI; BARBOSA; LINDEN, 2015).

Na estratégia desenvolvida (DIGIAMPIETRI; BARBOSA; LINDEN, 2015), o problema de desambiguação de nomes de autores foi tratado como um problema de classificação binária em inteligência artificial que visa a classificar cada par de referências de autor que potencialmente pertencem a um mesmo autor como "pertencentes a um mesmo autor" ou "não pertencentes a um mesmo autor". Para isto, inicialmente, as referências a autores são agrupadas em blocos e dentro de cada bloco todos pares de referências são comparados e catorze atributos/características são extraídos de cada par. Com base nestas características, um algoritmo de classificação é executado para tentar identificar se as duas referências efetivamente se referem ao mesmo autor. A seguir, a metodologia utilizada e alguns resultados são detalhados.

A metodologia utilizada para o desenvolvimento e teste da estratégia de desambiguação foi organizada nas seguintes atividades: revisão da literatura correlata (sumarizada na seção 2.6); obtenção dos dados; seleção da amostra; anotação manual da amostra; extração das características; teste, validação e análise dos resultados (DIGIAMPIETRI; BARBOSA; LINDEN, 2015).

Obtenção dos dados. Dois conjuntos de dados foram obtidos: toda a base do projeto DBLP foi copiada em novembro de 2014 e também foram copiadas as listas dos professores permanentes dos programas de pós-graduação em Ciência da Computação no Brasil da CAPES[7]. Estes últimos dados serão utilizados na atividade de seleção da amostra.

Seleção da amostra. A abordagem utilizada para a formação dos blocos foi criar um bloco para cada para "primeiro nome, sobrenome", agrupando os registros de autores do DBLP que possuíssem entre as diversas partes de seus nomes o primeiro nome e o último nome (sobrenome) que definem o bloco. A amostra utilizada para testes criou blocos a partir dos nomes dos 48 professores permanentes do programa de pós-graduação em Ciência da Computação da UNICAMP. A escolha deste programa foi feita arbitrariamente entre os programas nota 7 da área de Ciência da Computação. Ao se analisar os cerca de 1,5 milhão de registros de autores do DBLP, 82 registros fizeram parte dos 48 blocos criados,

[7] http://www.capes.gov.br/component/content/article?id=4656:ciencia-da-computacao

com a identificação única e correta de 29 professores (blocos com um único registro) e a criação de 17 blocos ambíguos contendo entre 2 e 12 registros de autores cada. Além disso, dois professores não tiveram nenhum registro encontrado no DBLP (blocos sem nenhum registro).

Anotação da amostra. Os 17 blocos formados com dois ou mais registros foram analisados manualmente de forma a se verificar se cada par de registros dentro do mesmo bloco corresponde ou não ao mesmo autor. Ao todo, 102 pares foram analisados e a anotação foi feita por uma pessoa e validada por outra (um aluno de mestrado e o autor do presente texto). A anotação foi utilizada para treinamento e teste da estratégia proposta.

Extração das características. A estratégia foi baseada na extração de 14 características de quatro tipos para cada um dos 102 pares de registros de autores pertencentes ao mesmo bloco. Os tipos de informação utilizados são: nomes dos autores; rede social de coautorias (criada a partir das coautorias em todas as publicações de artigos em periódicos ou anais de eventos da base DBLP); mineração de texto baseada nos títulos dos artigos (considerando os 1.828 artigos publicados pelos autores da amostra); e datas de publicação dos artigos dos autores da amostra (DIGIAMPIETRI; BARBOSA; LINDEN, 2015). A tabela 9 descreve cada uma das características extraídas.

Teste e análise dos resultados. Uma verificação da importância das características foi realizada em relação a sua influência na desambiguação de nomes, bem como a acurácia da estratégia em si foi verificada, conforme será detalhado a seguir.

A figura 5 contém os valores da correlação de Pearson entre todas as características, sendo que *classe* indica se um par de referência a autores corresponde (classe=1) ou não (classe=0) a um mesmo autor. A correlação entre a *classe* e as demais características pode ser observada na primeira linha de dados. As três maiores correlações (todas negativas) nesta linha ocorrem entre características relacionadas aos nomes dos autores: *proporção de diferentes nomes do meio*, *nomes ou abreviações diferentes* e *último nome diferente*. Este fato já era esperado para o problema de desambiguação de nomes e os maiores desafios para este problema ocorrem justamente quando este tipo de informação não é suficiente para concluir se duas referências se referem ou não ao mesmo autor. A característica não relacionada aos nomes dos autores que obteve a maior correlação com a *classe* foi *vizinhos em comum*, confirmando a hipótese de que duas referências potencialmente do mesmo autor que possuam vizinhos em comum (coautores) na rede têm maior chance de realmente se referirem a um único autor.

Tabela 9 – Características extraídas das citações

Tipo de característica	Característica	Descrição
características da rede social de coautorias	vizinhos em comum	número de vizinhos em comum na rede social acadêmica de todos os autores da DBLP
	são vizinhos	indica se o par de autores é ou não vizinho na rede social
características extraídas dos nomes	distância de edição	distância de edição entre os nomes dos dois autores
	distância relativa	proporção entre a distância de edição e o tamanho do menor nome dentre os autores
	primeiro nome diferente	indica se os autores têm seus primeiros nomes diferentes (um do outro)
	último nome diferente	indica se os autores têm seus últimos nomes diferentes (um do outro)
	proporção de diferentes nomes do meio	proporção (em relação ao número total de nomes) de nomes diferentes entre os autores, contabilizando nomes adicionais como diferentes
	proporção de diferentes abreviações	proporção (em relação ao número total de nomes) de nomes abreviados diferentes entre os autores, contabilizando abreviações adicionais como diferentes
	nomes invertidos	indica se há inversão da posição das partes dos nomes entre os autores
	nomes ou abreviações diferentes	indica a proporção de nomes efetivamente diferentes entre os autores (sem contar a presença de nomes adicionais)
características baseadas na mineração de texto	mineração de texto dos títulos dos artigos	métrica baseada em TFIDF que compara a frequência das palavras dos títulos entre os dois autores e entre o corpus formado pelos títulos de todos os autores avaliados
	log(MT)	logaritmo do valor resultante da mineração de texto
características baseadas nos anos de publicação dos artigos	intersecção do período de publicação	interseção entre os períodos de publicação dos dois autores
	distância em anos entre publicações	distância mínima em anos entre as publicações dos dois autores (apenas se a interseção for igual a zero)

Fonte: Digiampietri, Barbosa e Linden (2015)

Destaca-se que muitas das características com maior correlação com a *classe* possuem alta correlação entre si e, por isso, potencialmente não irão contribuir muito para o processo de desambiguação.

A característica *são vizinhos* não aparece em nenhuma das avaliações a seguir, pois, para todos os pares da amostra, ela possui valor igual a zero (nenhum dos pares era vizinho na rede de coautorias).

Diferentes seletores de atributos[8] foram utilizados para se verificar quais subconjuntos de atributos agregam mais informação em relação ao atributo/característica *classe*. A tabela 10 contém o resultado da execução de alguns destes seletores, os quais ranqueiam cada um dos atributos selecionados por eles. Os atributos mais selecionados são: *proporção de diferentes nomes do meio*, *nomes ou abreviações diferentes*, *último nome diferente*, *mineração de texto*, *log(MT)* e *primeiro nome diferente*. Destaca-se que as únicas características não relacionadas à comparação de nomes que apareceram nesta lista foram as relacionadas

[8] Neste trabalho foram utilizadas os seletores de atributos disponíveis no arcabouço Weka

Figura 5 – Correlação entre as características detalhadas na tabela 9

	Classe	vizinhos em comum	distância de edição	distância relativa	primeiro nome diferente	último nome diferente	proporção de diferentes nomes do meio	proporção de diferentes abreviações	nomes invertidos	nomes ou abreviações diferentes	mineracao de texto	log(MT)	intersecção do período de publicação	distância em anos entre publicações
classe	1.000	0.320	-0.244	-0.304	-0.163	-0.464	-0.622	0.054	-0.066	-0.560	-0.033	0.173	0.159	-0.035
vizinhos em comum	0.320	1.000	-0.152	-0.166	-0.068	-0.192	-0.331	-0.088	0.133	-0.232	-0.019	0.096	0.100	-0.087
distância de edição	-0.244	-0.152	1.000	0.945	0.220	0.258	0.382	0.012	0.041	0.312	0.021	-0.007	-0.225	0.143
distância relativa	-0.304	-0.166	0.945	1.000	0.263	0.268	0.378	0.048	0.087	0.322	0.073	-0.002	-0.195	0.135
primeiro nome diferente	-0.163	-0.068	0.220	0.263	1.000	0.081	0.317	-0.142	0.652	0.273	-0.083	-0.131	0.013	0.005
último nome diferente	-0.464	-0.192	0.258	0.268	0.081	1.000	0.416	0.022	0.235	0.418	0.045	-0.159	-0.086	0.019
proporção de diferentes nomes do meio	-0.622	-0.331	0.382	0.378	0.317	0.416	1.000	-0.357	0.130	0.669	0.096	-0.057	-0.186	0.116
proporção de diferentes abreviações	0.054	-0.088	0.012	0.048	-0.142	0.022	-0.357	1.000	-0.142	0.092	-0.036	-0.041	0.085	-0.127
nomes invertidos	-0.066	0.133	0.041	0.087	0.652	0.235	0.130	-0.142	1.000	0.020	-0.065	-0.101	-0.075	-0.002
nomes ou abreviações diferentes	-0.560	-0.232	0.312	0.322	0.273	0.418	0.669	0.092	0.020	1.000	0.099	-0.008	-0.142	0.097
mineracao de texto	-0.033	-0.019	0.021	0.073	-0.083	0.045	0.096	-0.036	-0.065	0.099	1.000	0.741	0.172	-0.110
log(MT)	0.173	0.096	-0.007	-0.002	-0.131	-0.159	-0.057	-0.041	-0.101	-0.008	0.741	1.000	0.341	-0.227
intersecção do período de publicação	0.159	0.100	-0.225	-0.195	0.013	-0.086	-0.186	0.085	-0.075	-0.142	0.172	0.341	1.000	-0.668
distância em anos entre publicações	-0.035	-0.087	0.143	0.135	0.005	0.019	0.116	-0.127	-0.002	0.097	-0.110	-0.227	-0.668	1.000

Fonte: adaptado de Digiampietri, Barbosa e Linden (2015)

à mineração dos títulos das publicações dos autores (DIGIAMPIETRI; BARBOSA; LINDEN, 2015).

Tabela 10 – Características ranqueadas por seletores de atributos

característica / seletor de atributos	GainRatio AttributeEval	CfsSubsetEval	ChiSquared AttributeEval	Filtered AttributeEval	InfoGain AttributeEval	SymmetricalUncert AttributeEval	ReliefF AttributeEval	SVMAttributeEval
vizinhos em comum							11	10
distância de edição							7	11
distância relativa							9	7
primeiro nome diferente	5		5	5	5	5	5	9
último nome diferente	3	1	3	3	3	3	1	3
proporção de diferentes nomes do meio	1	2	1	1	1	1	3	1
proporção de diferentes abreviações							8	8
nomes invertidos							13	12
nomes ou abreviações diferentes	2	3	2	2	2	2	2	2
mineração de texto	4	4	4	4	4	4	12	13
log(MT)	4		4	4	4	4	6	4
intersecção do período de publicação							4	5
distância em anos entre publicações							10	6

Fonte: Digiampietri, Barbosa e Linden (2015)

A classificação dos pares de referências como "pertencentes a um mesmo autor" ou "não pertencentes a um mesmo autor" foi realizada utilizando-se o metaclassificador *Rotation*

Forest (RODRIGUEZ; KUNCHEVA; ALONSO, 2006) e considerando as 14 características. A escolha do uso deste metaclassificador foi feita devido a bons resultados obtidos em alguns problemas similares. Foi utilizada a validação cruzada em 10 subconjuntos (*10-fold-crossvalidation*) para se mediar o desempenho do experimento realizado.

A tabela 11 contém o resultado do desempenho, avaliado segundo as medidas: Taxa de Verdadeiro-Positivos (VP); Taxa de Falso-Positivos (FP); Precisão; Revocação; Medida-F; e Área ROC (DIGIAMPIETRI; BARBOSA; LINDEN, 2015). Destaca-se que o sistema não classificou nenhum par erroneamente como "pertencentes a um mesmo autor" e foi capaz de identificar dois terços dos pares de referências pertencentes a um mesmo autor (ressalta-se que estes pares não haviam sido desambiguados corretamente pelos responsáveis pelo projeto DBLP, pois este experimento lidou justamente com a desambiguação das referências que estão cadastradas como diferentes no DBLP [não pertencentes a um mesmo autor]).

Tabela 11 – Desempenho da estratégia utilizada

classe	VP	FP	Precisão	Revocação	Medida-F	Área ROC
F	1	0,333	0,957	1	0,978	0,977
T	0,667	0	1	0,667	0,8	0,977
Média Ponderada	**0,961**	**0,294**	**0,962**	**0,961**	**0,957**	**0,977**

Fonte: Digiampietri, Barbosa e Linden (2015)

Uma verificação detalhada dos erros de classificação (isto é, pares de referências a um mesmo autor que não foram devidamente classificadas como "pertencentes a um mesmo autor") mostrou que todos os erros ocorreram quando um dos nomes do par a ser comparado era composto apenas por dois nomes (primeiro nome e último sobrenome) e não havia nenhuma outra característica (dentre as utilizadas) que evidenciasse que o par se refere a um mesmo autor.

Adicionalmente, foram realizados testes utilizando-se apenas subconjuntos das características/atributos. O menor subconjunto que obteve os mesmos resultados do uso de todos os atributos foi composto por quatro atributos: primeiro nome diferente, último nome diferente, proporção de diferentes nomes do meio, e mineração de textos dos títulos.

Considerou-se que os resultados iniciais foram bastante promissores indicando que a estratégia desenvolvida tem potencial para aprimorar o processo de desambiguação utilizado no projeto DBLP.

A presente estratégia de desambiguação de nomes ainda está em fase de aprimoramento. Pretende-se realizar testes utilizando maiores conjuntos de dados além de implementar extratores de outras características a serem utilizadas na classificação.

5 Análise de grupos

Este capítulo apresenta os dois principais tipos de análise de grupos que foram realizados pelo autor e seus colaborados ao longo dos últimos anos.

O primeiro teve por objetivo comparar diferentes características bibliométricas e de análise de redes sociais de forma a analisar o comportamento dos diferentes programas de pós-graduação na área de Ciência da Computação no Brasil de acordo com cada uma das características medidas e comparar este comportamento com a nota atribuída pela CAPES para os respectivos programas durante dois triênios.

O segundo teve por objetivo analisar grupos de pesquisadores considerando sua distribuição geográfica no país. Estes grupos foram organizados de acordo com suas áreas de atuação e/ou titulação.

As próximas seções detalham as análises realizadas, descrevendo as amostras utilizadas, alguns aspectos metodológicos, os resultados e conclusões.

5.1 Análise dos programas de pós-graduação em Ciência da Computação

Esta seção apresenta os principais resultados obtidos com a análise dos programas de pós-graduação em Ciência da Computação, especialmente os já publicados em Digiampietri et al. (2014).

Conforme apresentado, a análise e classificação ou ranqueamento de grupos é uma atividade complexa envolvendo alguns aspectos subjetivos e/ou de difícil mensuração. Adicionalmente, diferentes trabalhos da literatura correlata usam índices diversos ou os mesmos índices, mas com fontes de dados diferentes para a análise bibliométrica. Além de apresentar uma metodologia para a obtenção e organização dos dados, a pesquisa apresentada a seguir visou a caracterizar os programas de acordo com diferentes métricas e também a mostrar a diferença de ranqueamento que ocorre de acordo com a métrica escolhida e/ou a fonte de informação (no caso específico dos índices calculados a partir de citações). Destaca-se que não foi objetivo desta pesquisa propor nenhum novo tipo de avaliação para os programas brasileiros de pós-graduação e sim fazer uma caracterização e análise comparativa segundo algumas métricas (DIGIAMPIETRI et al., 2014).

Assim, o objetivo desta pesquisa foi caracterizar os programas brasileiros de pós-graduação em Ciência da Computação e os relacionamentos de coautoria entre eles. Para isto, os seguintes objetivos específicos foram traçados: (i) quantificar diferentes características dos programas; (ii) ranquear os programas de acordo come estas características; (iii) analisar a correlação entre as características; (iv) analisar o relacionamento dos programas por meio da análise da rede de coautorias. Dois tipos de métricas foram utilizados: bibliométricas partindo-se das informações dos Currículos Lattes dos pesquisadores e enriquecidas com informações acerca das citações (oriundas dos sites Google Scholar e Microsoft Academic Search) e dos veículos de publicação (índice JCR: *Thompson's Journal Citation Reports* e SJR: *Scimago Journal Rank* e Qualis); e métricas relacionadas à análise de redes sociais, em particular medidas de centralidade e aglomeração.

5.1.1 Materiais e métodos

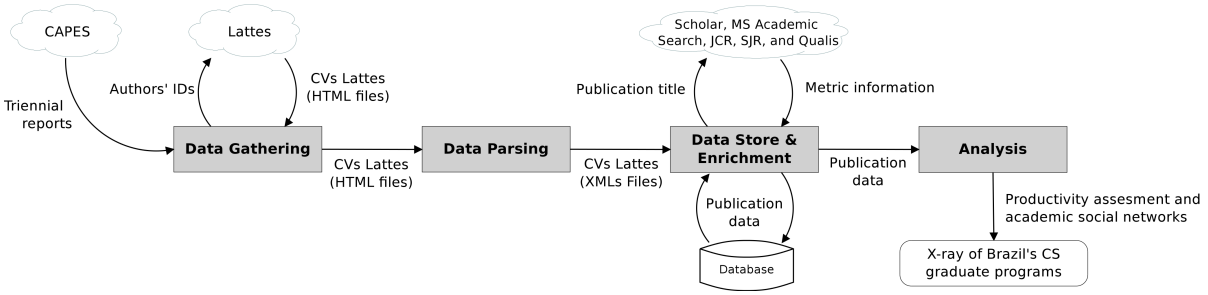
Foram examinados os 37 programas brasileiros de pós-graduação em Ciência da Computação que possuíam mestrado acadêmico e/ou doutorado em ambos os triênios 2004-2006 e 2007-2009. A identificação destes programas e de seus docentes permanentes foi extraída dos *Cadernos de Indicadores* da CAPES[1]. Com base nestes documentos foram encontrados os identificadores dos currículos dos 732 professores permanentes destes programas e seus currículos foram copiados da Web no formato HTML. Estes currículos foram processados conforme apresentado no capítulo 3 e dois bancos relacionais foram criados, um para cada triênio.

Uma das informações extraídas dos currículos mais importantes para as análises realizadas é referente aos artigos completos publicados em periódicos ou anais de eventos, pois são a base para a análise bibliométrica e análise da rede de coautorias. Ao todo, 17.976 publicações foram encontradas (13.926 artigos completos em conferências e 4.050 artigos publicados em periódicos). A identificação de coautorias utilizou, além do título dos artigos, a lista de autores, ano de publicação, veículo de publicação, páginas e volume, conforme metodologia publica por Digiampietri et al. (2012b) e sumarizada na seção 4.1. Ao todo foram identificadas 1.428 relações de coautorias entre pares de pesquisadores pertencentes a amostra utilizada.

[1] www.capes.gov.br/avaliacao/documentos-de-area-/3270

A figura 6 apresenta o fluxo de atividades utilizado para as análises desenvolvidas nesta parte da pesquisa (DIGIAMPIETRI et al., 2014). Nesta figura, as nuvens representam fontes de dados, os retângulos representam as atividades desenvolvidas e as setas representam fluxos de informação.

Figura 6 – Diagrama das atividades desenvolvidas



Fonte: Digiampietri et al. (2014)

A primeira informação adicionada aos bancos relacionais de Currículos Lattes criados foi a atribuição de cada professor ao programa de pós-graduação em que atua e o período de atuação. Esta última informação foi necessária, pois nem todos os professores atuaram durante todo o triênio. Estas informações foram extraídas a partir dos *Cadernos de Indicadores* da CAPES.

As demais informações adicionadas na etapa de *Enriquecimento de Dados* são informações bibliométricas. Foram adicionadas a cada publicação: fator de impacto JCR (*Thompson's Journal Citation Reports*), índice SJR (*Scopus's Scimago Journal Rank*) e classificação Qualis de cada veículo de publicação; citações obtidas de duas fontes: *Microsoft Academic Search* e *Google Scholar*. Adicionalmente, os índices derivados das citações foram calculados para cada um dos programas considerando todas as publicações de seus docentes no respectivo triênio: índice h e índice g, calculados com base nas citações das duas fontes utilizadas.

Para converter as classificações Qualis para valores numéricos foram utilizados os pesos presentes no documento de área da Ciência da Computação (a saber: A1=100, A2=85, B1=70, B2=50, B3=20, B4=10, B5=5 e C=0).

Após a realização da obtenção, organização e enriquecimento dos dados, dois tipos de análise foram desenvolvidas: análise bibliométrica e análise da rede social de coautorias. As análises foram realizadas considerando os programas como unidades básicas e as métricas foram extraídas do conjunto das publicações dos professores de cada um

dos programas e as coautorias entre programas refletem as coautorias entre docentes de diferentes programas.

5.1.2 Resultados

Os programas foram ranqueados de acordo com dez métricas (sendo as cinco primeiras e a última ponderadas pelo número de docentes permanentes em cada programa): (i) citações utilizando dados do Microsoft Academic Search (*MS CC*), (ii) citações utilizando dados do Google Scholar (*MS CC*); (iii) soma dos valores do fator de impacto JCR dos periódicos de todas as publicações de cada programa dividida pelo número de professores permanentes do programa (*JCR*); (iv) *SJR* e (v) *Qualis*, correspondendo, respectivamente a soma dos índices SJR ou Qualis dos veículos de publicação de todas as publicações de cada programa dividida pelo número de professores permanentes do programa; duas versão do índice-h e do índice-g variando apenas de acordo com a fonte de citações utilizada: (vi) *MS h-index*, (vii) *MS g-index*, (viii) *GS h-index*, e (ix) *GS g-index*; e (x) número total de publicações dividido pelo número de professores permanentes (*Pubs. count*). A tabela 12 apresenta a classificação de cada programa segundo estas métricas e, adicionalmente, a melhor classificação, a pior e a mediana. Os dados desta tabela consideraram os artigos completos publicados em revistas e anais de eventos no período de 2004 a 2009 (DIGIAMPIETRI et al., 2014).

Observa-se a grande variação de classificação dos programas dependendo da métrica utilizada, lembrando-se que nenhuma, por si só, é capaz de avaliar de maneira completa e robusta todos os aspectos envolvidos e desejáveis de um programa de pós-graduação.

A figura 7 sumariza de maneira gráfica as informações da tabela 12 por meio de um *box-plot*. Nesta figura os programas foram ordenados de acordo com a mediana dos valores de seus ranqueamentos. Por exemplo, o programa indicado como 2 teve seus ranqueamentos variando de 1 a 7, sendo a mediana de seus ranqueamentos igual a 2,5 (tabela 12 e figura 7).

A figura 8 apresenta a dinâmica da variação dos ranqueamentos entre o triênio 2004-2006 e o triênio 2007-2009. É possível observar que, na mediana, há programas que subiram mais de dez posições considerando as métricas analisadas (como é o caso dos dois primeiros programas apresentados na figura 8, identificados como programas 17 e 29), vários permaneceram, na mediana, estáveis (programas 18, 2, 8, 9, 37 e 5); e outros tiveram

Tabela 12 – Programas brasileiros de pós-graduação em Ciência da Computação ranqueados de acordo com diferentes métricas

CAPES **P**rog. #	**MS** **C**C	**S**cholar **C**C	**J**CR	**S**JR	**Q**ualis	MS **h**-index	MS **g**-index	**S**cholar **h**-index	**S**cholar **g**-index	**P**ubs. **c**ount	**B**est **r**ank	**W**orst **r**ank	**M**edian **r**ank
1	2º	2º	13º	11º	10º	5º	6º	6º	5º	6º	2º	13º	6º
2	1º	1º	2º	1º	4º	3º	2º	3º	3º	7º	1º	7º	2.5º
3	5º	6º	4º	2º	5º	22º	26º	20º	26º	4º	2º	26º	5.5º
4	6º	7º	10º	5º	2º	27º	28º	31º	32º	5º	2º	32º	8.5º
5	3º	3º	5º	3º	6º	23º	24º	26º	27º	2º	2º	27º	5.5º
6	4º	4º	3º	4º	8º	17º	15º	18º	19º	12º	3º	19º	10º
7	18º	11º	7º	6º	7º	37º	36º	35º	36º	9º	6º	37º	14.5º
8	11º	8º	8º	15º	12º	14º	17º	13º	15º	13º	8º	17º	13º
9	14º	15º	1º	13º	3º	24º	22º	27º	23º	34º	1º	34º	18.5º
10	13º	17º	12º	7º	9º	9º	12º	11º	13º	14º	7º	17º	12º
11	12º	12º	14º	9º	14º	8º	14º	9º	14º	17º	8º	17º	13º
12	10º	9º	9º	16º	13º	4º	4º	8º	4º	31º	4º	31º	9º
13	23º	24º	15º	17º	21º	25º	19º	22º	20º	32º	15º	32º	21.5º
14	9º	10º	25º	18º	17º	7º	7º	4º	7º	11º	4º	25º	9.5º
15	25º	25º	31º	33º	29º	28º	31º	24º	28º	28º	24º	33º	28º
16	27º	33º	17º	29º	30º	19º	16º	29º	25º	36º	16º	36º	28º
17	26º	26º	21º	23º	22º	32º	27º	34º	31º	26º	21º	34º	26º
18	19º	21º	11º	10º	18º	31º	32º	32º	35º	21º	10º	35º	21º
19	17º	16º	24º	8º	11º	15º	18º	15º	16º	19º	8º	24º	16º
20	28º	27º	16º	22º	25º	35º	37º	37º	37º	18º	16º	37º	27.5º
21	21º	19º	27º	24º	23º	21º	29º	21º	24º	16º	16º	29º	22º
22	34º	34º	26º	28º	34º	30º	34º	28º	33º	33º	26º	34º	33º
23	24º	28º	22º	26º	26º	13º	11º	19º	18º	25º	11º	28º	23º
24	20º	20º	20º	20º	19º	12º	21º	10º	21º	10º	10º	21º	20º
25	7º	5º	6º	12º	1º	2º	5º	2º	2º	3º	1º	12º	4º
26	16º	18º	23º	21º	20º	10º	8º	14º	10º	20º	8º	23º	17º
27	29º	31º	18º	25º	15º	6º	9º	7º	9º	22º	6º	31º	16.5º
28	22º	22º	19º	19º	16º	11º	10º	5º	8º	29º	5º	29º	17.5º
29	15º	14º	30º	14º	24º	1º	1º	1º	1º	1º	1º	30º	7.5º
30	35º	35º	35º	31º	28º	29º	30º	30º	29º	23º	23º	35º	30º
31	32º	30º	28º	32º	36º	33º	23º	33º	22º	37º	22º	37º	32º
32	8º	13º	34º	30º	27º	20º	3º	25º	6º	8º	3º	34º	16.5º
33	30º	23º	37º	36º	31º	18º	25º	12º	17º	15º	12º	37º	24º
34	31º	32º	33º	27º	35º	16º	13º	17º	11º	24º	11º	35º	25.5º
35	36º	36º	32º	35º	33º	34º	33º	36º	34º	30º	30º	36º	34º
36	33º	29º	29º	34º	32º	26º	20º	16º	12º	27º	12º	34º	28º
37	37º	37º	36º	37º	37º	36º	35º	23º	30º	35º	23º	37º	36º

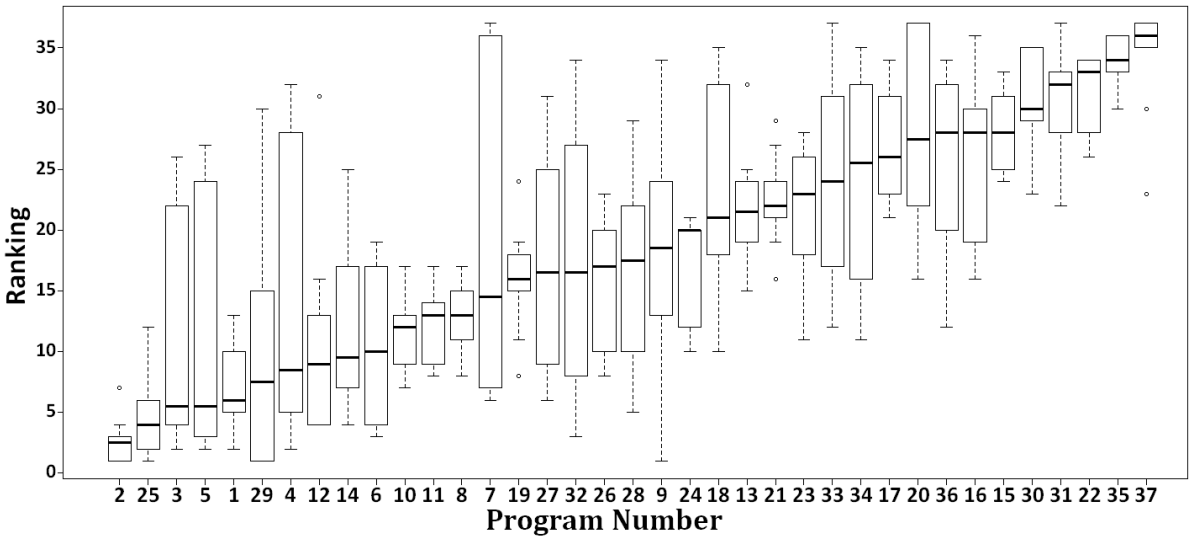
Fonte: Digiampietri et al. (2014)

variações negativas de, na mediana, mais de dez posições (por exemplo, os programas 14 e 25).

Duas medidas de correlação entre os ranqueamentos utilizando as diferentes métricas foram calculadas. A primeira corresponde à correlação de Spearman entre as classificações geradas por todos os pares de métricas (figura 9). A segunda corresponde à correlação entre cada um dos ranqueamentos e o valor da mediana dos ranqueamentos (figura 10) e visa a identificar qual dos ranqueamentos está mais correlacionado com a mediana dos mesmos.

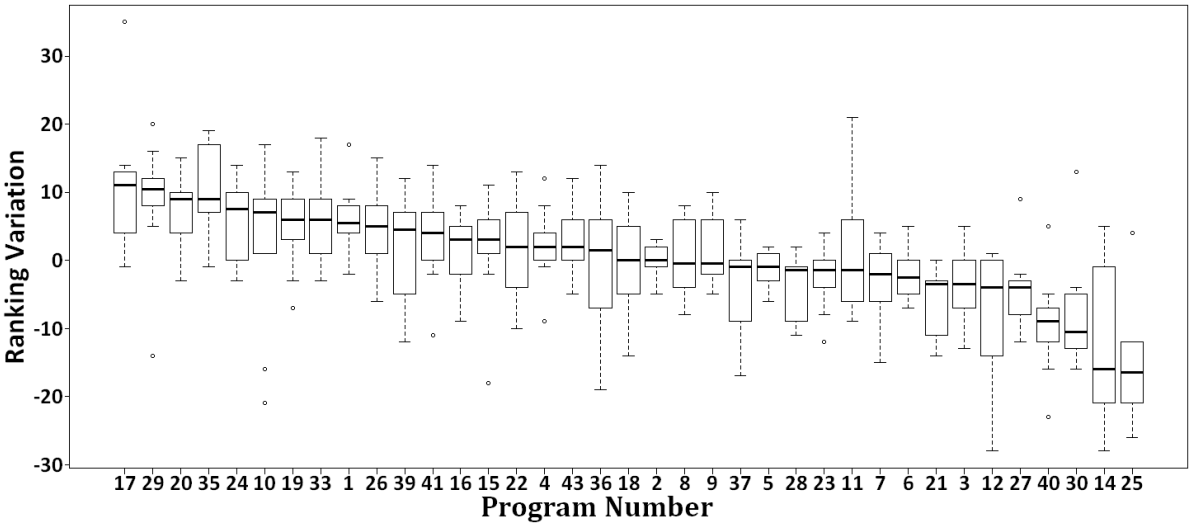
Na figura 9 é possível observar três grupos de ranqueamentos altamente correlacionados: (i) os quatro que utilizaram os índices h e g; (ii) os ranqueamentos utilizando JCR,

Figura 7 – Variação da classificação dos programas de acordo com a medida utilizada



Fonte: Digiampietri et al. (2014)

Figura 8 – Diferença de ranqueamento das diferentes métricas entre os triênios 2007-2009 e 2004-2006



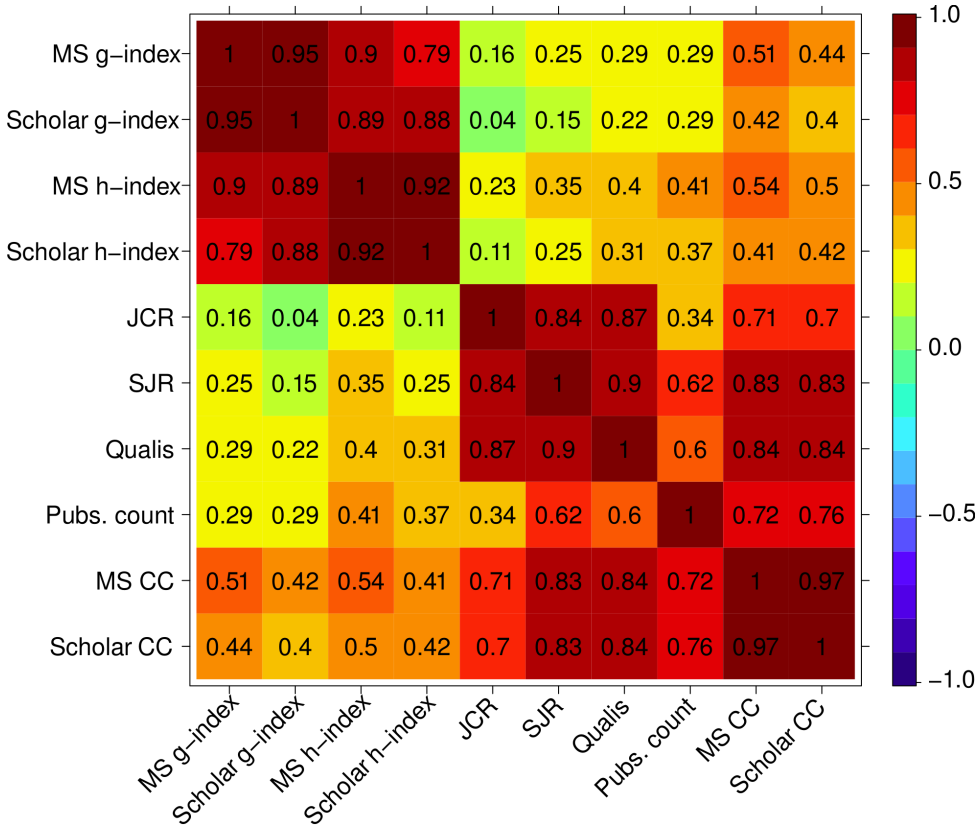
Fonte: Digiampietri et al. (2014)

SJR e Qualis; (iii) os dois ranqueamentos baseados na contagem de citações dividida pelo número de professores.

Pela figura 10 é possível observar que os dois ranqueamentos mais correlacionados com a mediana dos ranqueamentos são aqueles ligados com a contagem de citações. Em particular, a maior correlação foi obtida pelo ranqueamento utilizando as citações do Google Scholar (*Scholar CC*).

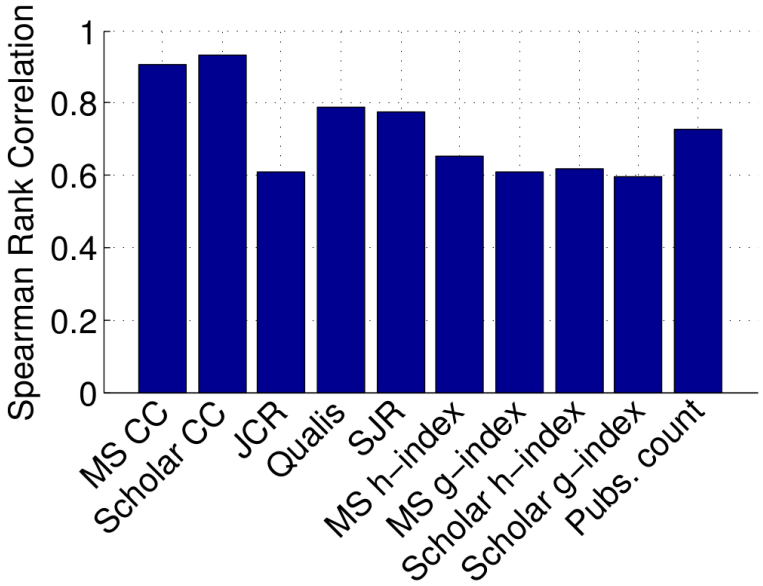
Estava fora do contexto deste trabalho uma comparação detalhada entre o ranqueamento utilizando as notas atribuídas pela CAPES nas avaliações trienais e os ranqueamentos

Figura 9 – Correlação de Spearman entre os diferentes ranqueamentos



Fonte: Digiampietri et al. (2014)

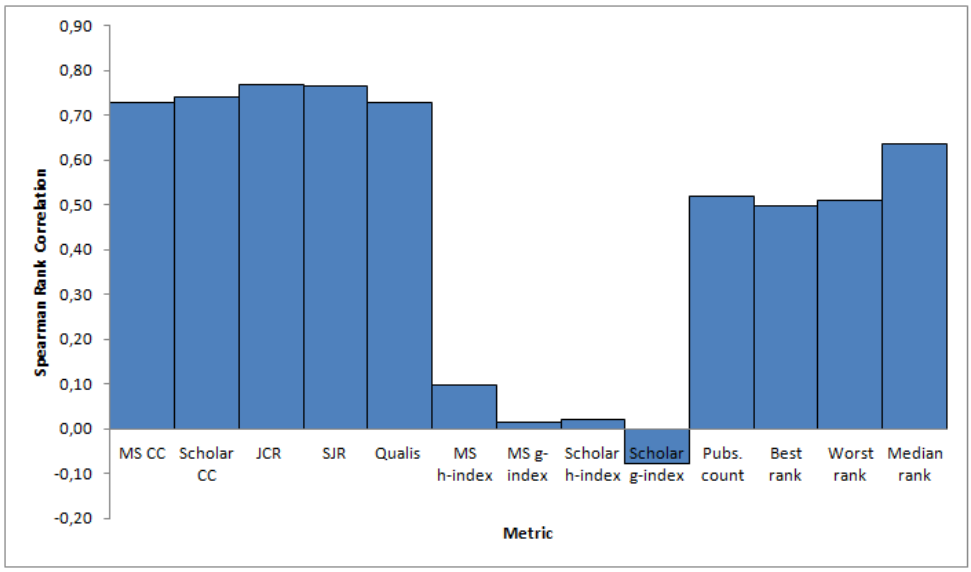
Figura 10 – Correlação de Spearman entre os valores dos diferentes ranqueamentos e a mediana dos mesmos



Fonte: Digiampietri et al. (2014)

utilizando as diferentes métricas calculadas neste trabalho. Porém, a título de curiosidade, a figura 11 apresenta a correlação entre o ranqueamento utilizando-se a nota CAPES e os ranqueamentos apresentados na tabela 12. Os cinco ranqueamentos mais correlacionados são os três que consideram índices de qualidade/impacto dos veículos de publicação e os dois que utilizam a contagem de citações.

Figura 11 – Correlação de Spearman entre os diferentes ranqueamentos e o ranqueamento usando a nota CAPES do triênio 2007-2009



Fonte: Digiampietri (2015)

Para a análise de redes sociais, duas redes de coautoria foram produzidas, cada uma utilizando dados de 2004 a 2009. A primeira é uma rede de coautorias não direcionada na qual haverá uma aresta entre o programa 1 e o programa 2 se ao menos um autor do programa 1 colaborou com ao menos um autor do programa 2 na publicação de um ou mais artigos no período. Assim, haverá uma aresta não direcionada ligando os programas 1 e 2. A segunda é uma rede direcionada na qual haverá uma aresta do programa 1 para o programa 2 se houver um artigo no período no qual um professor do programa 1 é o primeiro autor e um ou mais professores do programa 2 são coautores deste artigo. O relacionamento denotado por estas arestas direcionadas é as vezes chamado de relacionamento de "solicitação de ajuda" ou "convite para colaborar" e possibilita alguns estudos mais aprofundados das redes de coautoria do que aqueles utilizando o relacionamento não direcionado de colaboração em publicações (NEWMAN, 2004).

As métricas utilizadas para caracterizar cada um dos nós (programas) nas redes de coautoria já foram apresentadas na seção 2.1, mas serão brevemente sumarizadas a seguir.

Três métricas de centralidade foram calculadas: de grau (*Cnt.Deg*), de intermediação (*Cnt.Bet*), e de proximidade (*Cnt.Clos*) (BONACICH, 1987). Medidas de centralidade objetivam identificar quais são os elementos mais importantes/centrais em uma rede. O coeficiente de agrupamento/*clusterização* costuma ser utilizado para medir o quão transitivas são as relações de coautoria, isto é, se o programa 1 está ligado ao programa 2 e o programa 2 está ligado ao programa 3, o coeficiente de agrupamento irá medir a probabilidade do programa 1 também estar ligado ao programa 3. Dentro de uma rede de coautorias na qual cada autor é um nó, um alto coeficiente de clusterização costuma indicar uma coesão no grupo, porém na rede na qual os programas são os nós, um coeficiente de agrupamento mais baixo pode indicar que o programa exerce uma função de ligação importante entre outros programas e, assim, um valor mais baixo para esta medida pode indicar um papel importante do programa no contexto nacional (MELO; ALMEIDA; LOUREIRO, 2008).

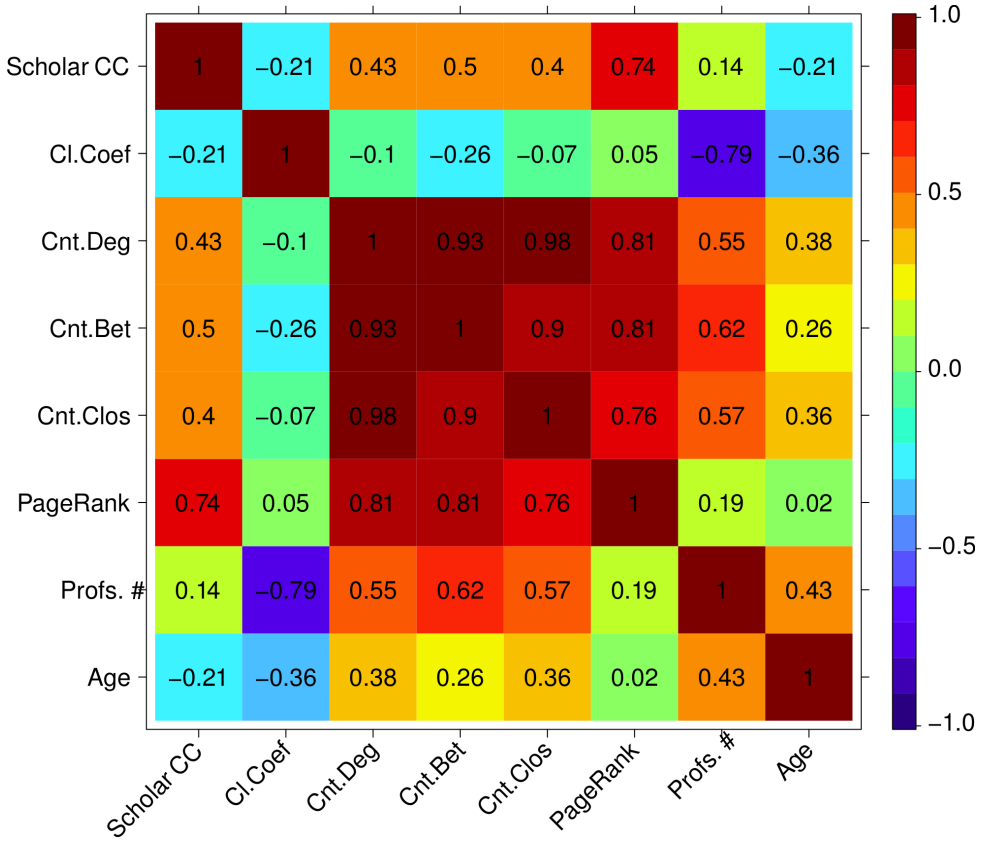
As métricas apresentadas foram calculadas considerando a rede não direcionada. Para a rede direcionada, uma única métrica relacionada a centralidade foi calculada: o *Page Rank*. Esta métrica é calculada pelo algoritmo de ranqueamento de páginas Web desenvolvido pela empresa Google e visa a identificar nós importantes em uma rede direcionada. Para isto, o algoritmo, iterativamente, transfere parte da importância atribuída a cada nó para cada um dos nós para os quais este nó aponta (LANGVILLE; MEYER, 2009). No caso da rede direcionada criada, o algoritmo deverá ser capaz de identificar quais programas foram mais convidados a participar de colaborações em publicações.

Estas métricas foram computadas para cada um dos programas e foram calculadas as correlações entre estas métricas par a par e entre cada uma dessas e a contagem de citações usando dados do Google Scholar dividida pelo número de docentes de cada programa (*Scholar CC*). Optou-se por só utilizar esta medida bibliométrica para o cálculo das correlações pelo fato de ela ter apresentado a maior correlação com as demais medidas bibliométricas utilizadas.

A figura 12 apresenta as correlações. Destaca-se que a maior correlação entre as citações (*Scholar CC*) e as métricas de redes ocorre com a métrica *PageRank*. Adicionalmente, é possível observar que as quatro métricas de centralidade são altamente correlacionadas.

As figuras 13 e 14 apresentam, respectivamente, as redes de coautoria não-direcionada e direcionada. Nestas redes, os programas estão coloridos de acordo com a região geográfica em que os programas estão no Brasil (por exemplo, a cor verde corresponde à região

Figura 12 – Correlação de Spearman entre as diferentes métricas da rede



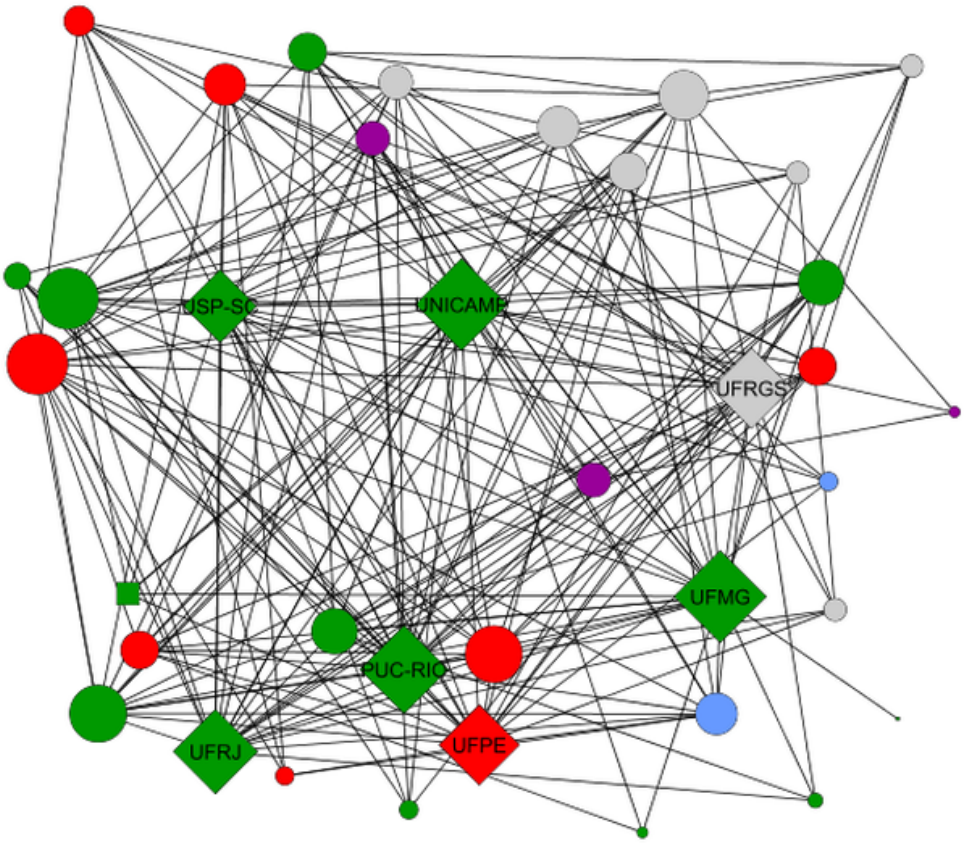
Fonte: Digiampietri et al. (2014)

Sudeste e a cor cinza à região Sul). Os programas que no triênio 2007-2009 possuíam maior nota CAPES estão rotulados nas redes. As redes/grafos foram desenhados utilizando o algoritmo (FDLA) (FRUCHTERMAN; REINGOLD, 1991), que visa a minimizar a quantidade de intersecções entre arestas. O tamanho dos nós em cada uma das redes é proporcional ao grau do nó (rede não-direcionada) e ao valor de seu *PageRank* (rede direcionada).

É possível observar que, nas duas redes, a maioria dos nós com maior tamanho (maior grau ou maior valor de *PageRank*) corresponde justamente aos programas com maior nota atribuída pela CAPES (programas com os rótulos nos nós).

A fim de visualizar de maneira mais prática os programas de acordo com as métricas de redes calculadas foi utilizada a análise de componentes principais (*Principal Component Analysis (PCA)* (JOLLIFFE, 2002) para reduzir a dimensionalidade dos dados de forma a possibilitar a criação de uma figura bidimensional que ilustrasse a posição dos programas. Para isto, foram utilizadas as duas componentes principais. A análise de componentes principais é uma técnica estatística muito usada para a redução de dimensionalidade de forma não supervisionada. Esta técnica cria uma transformação dos dados originais em um

Figura 13 – Rede de coautoria não-direcionada dos programas de pós-graduação em Ciência da Computação

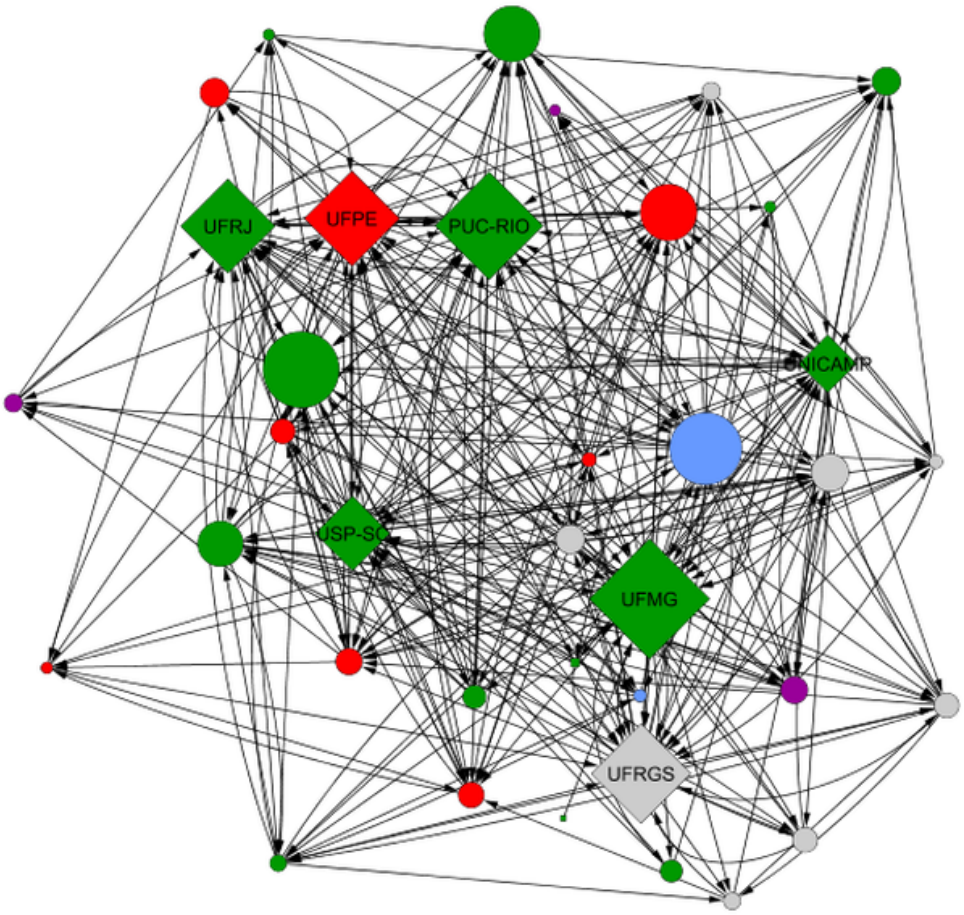


Fonte: Digiampietri et al. (2014)

novo conjunto de características/atributos chamados componentes principais. O primeiro componente é aquele que possui a maior variância possível, o segundo possui a maior variância possível (mas já considerando a variância do primeiro, sendo assim ortogonal ao primeiro) e assim por diante (JOLLIFFE, 2002). Para o conjunto de características de redes, as duas primeiras componentes principais foram capazes de representar quase 95% da variância. A primeira componente principal foi capaz de representar 77% da variância e a segunda 18%.

A figura 15 apresenta o gráfico dos programas de acordo com as duas primeiras componentes principais, a primeira está no eixo da abscissa e a segunda no eixo da ordenada. Nesta figura é possível observar como ocorreu o mapeamento de cada uma das características. Nota-se que a primeira componente está mais relacionada às medidas de centralidade e, a segunda, ao coeficiente de agrupamento. É interessante notar que todos os programas com notas 6 ou 7 estão agrupados do lado direita da figura (representados por losangos amarelos), já na parte central da figura estão todos os programas nota 5 e

Figura 14 – Rede de coautoria direcionada dos programas de pós-graduação em Ciência da Computação



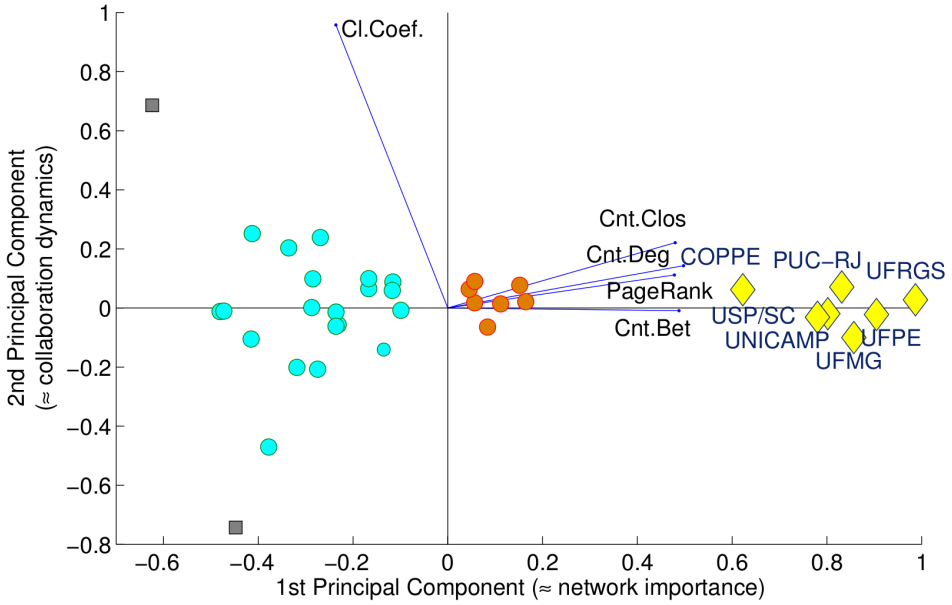
Fonte: Digiampietri et al. (2014)

alguns nota 4. Destaca-se ainda que seria possível separar os programas notas 6 e 7 dos demais utilizando-se apenas a primeira componente principal.

A maioria dos resultados apresentados nesta seção foram publicados em (DIGIAMPIETRI et al., 2014). Neste artigo também foi realizada uma análise minuciosa sobre a dinâmica das arestas entre os programas de pós-graduação em Ciência da Computação considerando os triênio 2004-2006 e 2007-2009

Dentro deste contexto, outros estudos foram realizados especificamente sobre a dinâmica das redes de coautorias dos professores dos programas de pós-graduação em Ciência da Computação, nos quais cada nó da rede era um professor ao invés de um programa (DIGIAMPIETRI et al., 2012b; DIGIAMPIETRI et al., 2015b). A figura 16 ilustra as colaborações históricas dos professores permanentes dos programas de pós-graduação em Ciência da Computação que estavam credenciados no triênio 2007-2009. O tamanho dos nós é proporcional a medida *Author Rank* (LIU et al., 2005) de cada professor e os nós

Figura 15 – Gráfico dos programas de pós-graduação em Ciência da Computação considerando as duas primeiras componentes principais

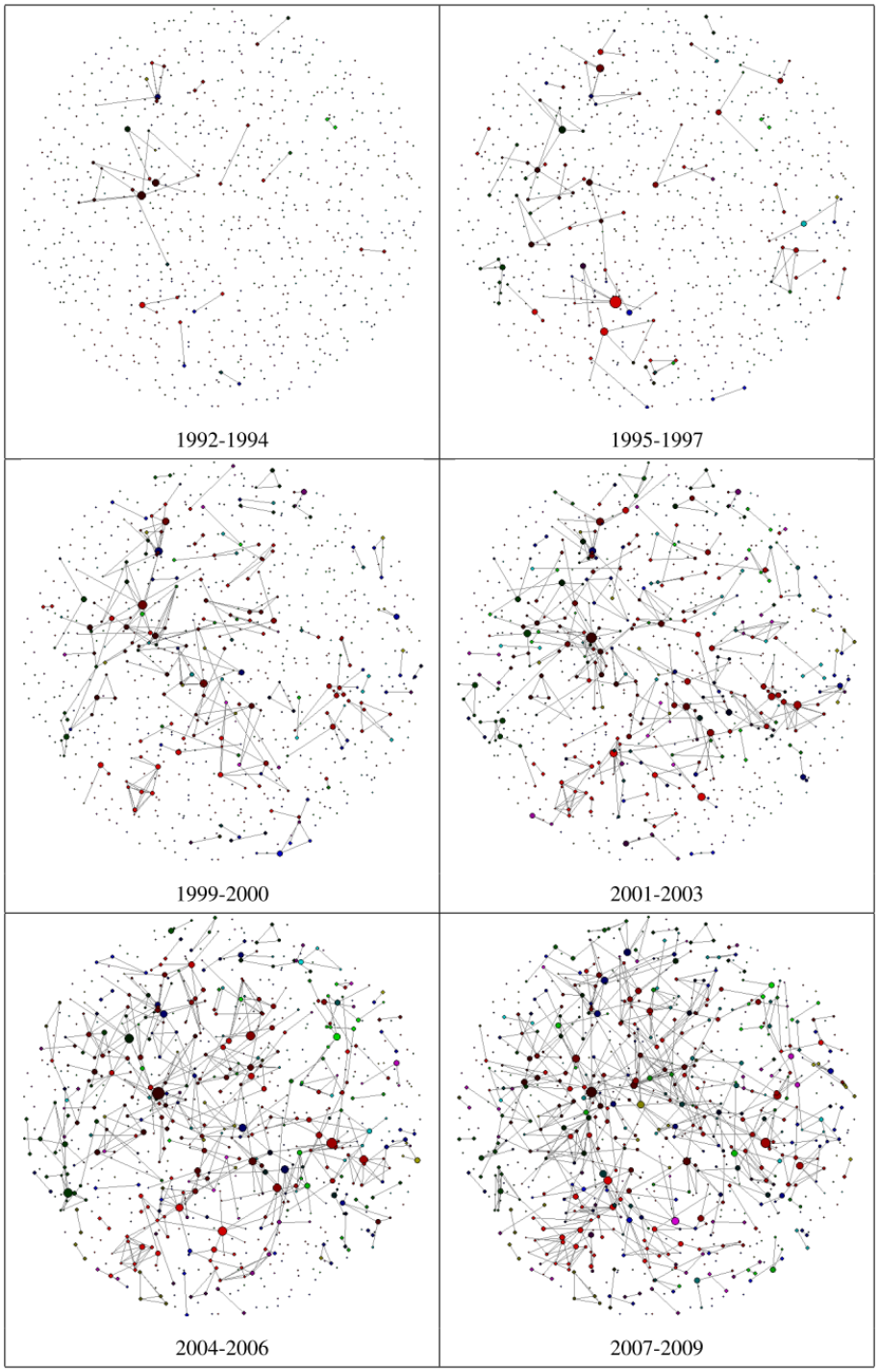


Fonte: Digiampietri et al. (2014)

foram coloridos de acordo com o programa em que os professores estão credenciados. Esta dinâmica também foi explorada na pesquisa sobre predição de relacionamentos que será apresentada no capítulo 6.

A figura 17 apresenta a evolução no número de publicações ao longo dos anos considerando os professores permanentes que estavam credenciados nos programas de pós-graduação em Ciência da Computação no triênio 2007-2009. Já a figura 18 apresenta a relação entre o número de publicações em coautoria e o número total de publicações diferentes. Observa-se ao longo dos triênios um aumento da produção deste conjunto de docentes e, em especial, um aumento relativo das publicações em coautoria. Os dados utilizados nestas figuras foram obtidos em 2011 e, provavelmente, alguns dos currículos não estavam devidamente atualizados (DIGIAMPIETRI et al., 2014b), isto pode explicar a aparente diminuição da produção e da colaboração no último triênio analisado em relação ao anterior. De fato, em um estudo em andamento utilizando os mesmos professores foi observado que no triênio 2007-2009 a produção foi superior aquela do triênio anterior.

Figura 16 – Redes de coautorias entre os docentes dos programas de pós-graduação em Ciência da Computação

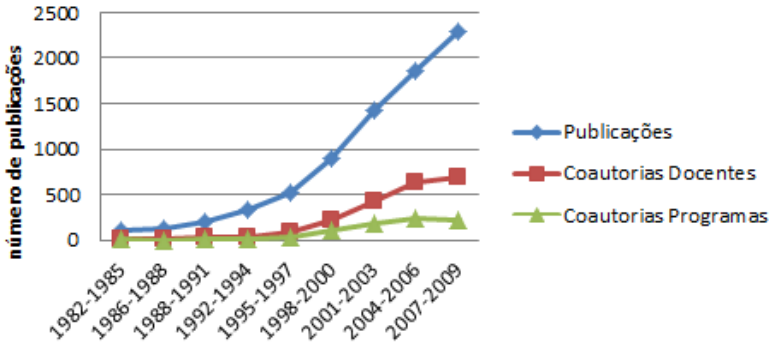


Fonte: Digiampietri et al. (2012)

5.1.3 Conclusões

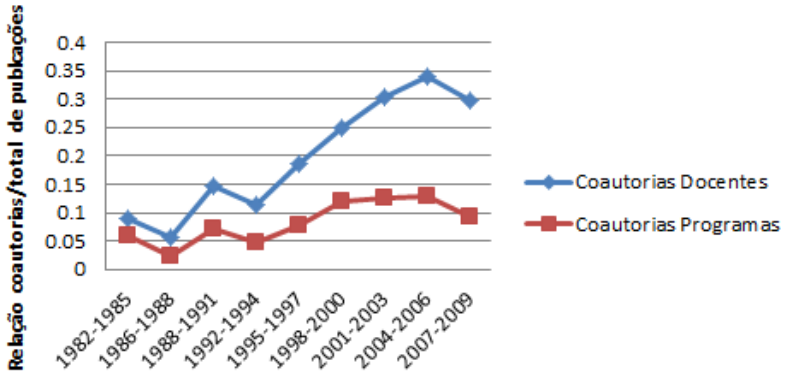
Neste estudo os programas de pós-graduação em Ciência da Computação foram analisados considerando-se diferentes medidas tanto bibliométricas quando oriundas da

Figura 17 – Evolução no número de publicações



Fonte: Digiampietri et al. (2012)

Figura 18 – Evolução entre coautorias e o número total de publicações



Fonte: Digiampietri et al. (2012)

análise de redes sociais. Foi possível observar as principais correlações entre as diferentes medidas e entre as mesmas métricas utilizando-se dados de diferentes fontes.

Um dos resultados mais interessantes está relacionado à constatação que os programas com melhor nota foram aqueles que se destacaram na análise de componentes principais realizada, formando um grupo separado dos demais.

Ao se analisar a dinâmica das redes entre os dois triênios (2004-2006 e 2007-2009) ou mesmo englobando triênios anteriores é possível observar um fortalecimento gradual nas coautorias tanto entre docentes quanto entre programas. É possível também identificar que alguns dos elementos mais centrais da rede são justamente os indivíduos que ligam diferentes programas. Esta informação pode indicar que uma boa política para o fortalecimento dos programas de pós-graduação nacionais é incentivar projetos de professores visitantes entre os programas nacionais (DIGIAMPIETRI et al., 2012).

5.2 Análise de pesquisadores de acordo com sua distribuição geográfica

Outro recorte utilizado na presente pesquisa para a análise de grupos foi baseado na distribuição geográfica dos pesquisadores, na formação e nas áreas de atuação.

Devido à grande dimensão e diversidade brasileira, tanto no aspecto geográfico quanto social e cultural, a análise de qualquer característica nacional é desafiadora. Porém, para a criação de políticas científicas nacionais, é fundamental que se tenha um profundo entendimento das características regionais de forma a se potencializar a efetividade dessas políticas.

O estudo apresentado nesta seção objetivou aumentar o entendimento da rede social acadêmica brasileira por meio do estudo da produção e das ligações entre os doutores que atuam no Brasil e possuem currículo cadastrado na Plataforma Lattes. Para isto, foram estudadas as distribuições dos doutores pelos estados brasileiros e de acordo com suas áreas de atuação. Adicionalmente, uma análise sobre os títulos das publicações destes doutores também foi realizada (DIGIAMPIETRI et al., 2014a).

5.2.1 Metodologia

A pesquisa apresentada nesta seção (e cujos principais resultados já estão publicados em Digiampietri et al. (2014a)) foi dividida em quatro atividades: obtenção dos dados e seleção da amostra; seleção das informações de interesse; cálculo de métricas; e análise dos resultados.

Obtenção dos dados e seleção da amostra. Foram obtidos, em julho de 2013, 3,2 milhões de currículos da Plataforma Lattes utilizando a estratégia apresentada na seção 3.1. Destes, foram selecionados aqueles que atendessem a dois critérios: possuir doutorado (informação obtida dos dados de "formação/titulação"); e possuir endereço profissional no Brasil (informação obtida dos "dados gerais" de cada currículo). 156.278 currículos atenderam aos critérios.

Seleção das informações de interesse. As informações que foram consideradas relevantes para o estudo realizada são: endereço profissional; título dos artigos completos publicados em periódicos e em anais de eventos; ligações explícitas entre os currículos (as ligações foram utilizadas para a montagem das redes sociais), e áreas de atuação.

Cálculo de métricas. Foram utilizadas métricas oriundas da análise de redes sociais/teoria dos grafos (tabela 13) e métricas de mineração de textos, estas últimas para analisar algumas características dos títulos das publicações dos doutores. As análises foram realizadas considerando 28 conjuntos de dados: um composto pela rede social dos 156.278 doutores; 26 conjuntos correspondendo a cada um dos estados brasileiros; e um conjunto correspondendo aos dados relativos ao Distrito Federal.

Análise dos resultados. Os dados referentes aos 28 conjuntos foram analisados e comparados conforme será apresentado na próxima subseção.

Tabela 13 – Métricas oriundas da teoria dos grafos utilizadas

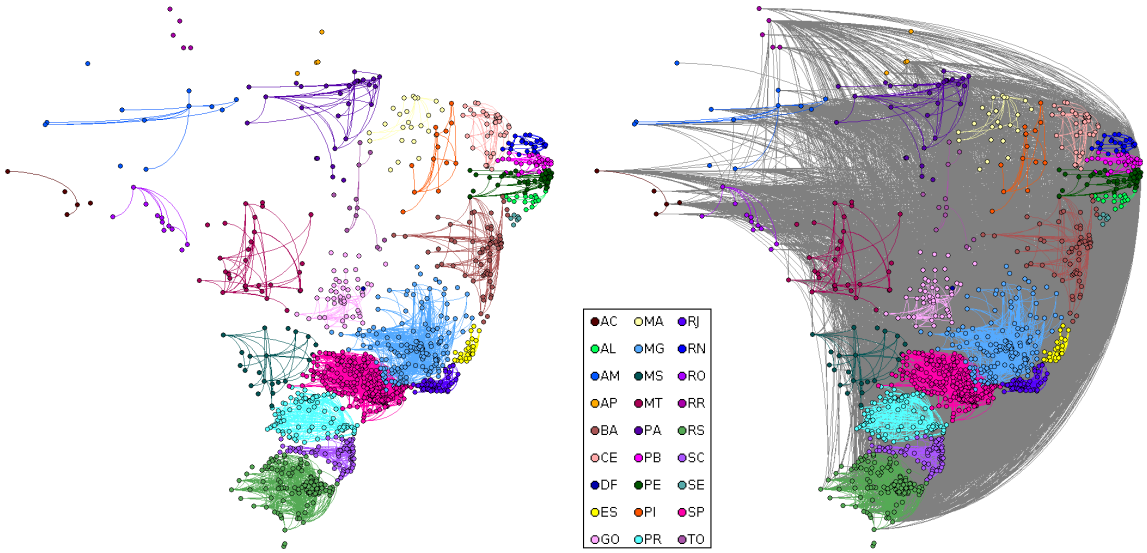
Métrica	Descrição
Nós	Número de nós presente na rede atual.
Arestas	Número de arestas presente na rede atual.
Nós no Componente Gigante	Número de nós no componente gigante (maior componete conexo).
Porcentagem de Nós no Componente Gigante	Porcentagem dos nós no componente gigante em relação a todos os nós da rede atual.
Densidade	Número de arestas do grafo atual em relação ao número máximo possível.
Grau Médio	Grau médio dos nós da rede atual.
Coeficiente de Clusterização	Quantidade de cliques de tamanho três dividida pela quantidade de três nós conectados.
Assortatividade de Grau	Métrica que calcula a tendência de nós de se conectarem a nós de mesmo grau (1 indica que todos os nós se conectam apenas a nós de mesmo grau e -1 indica que todos os nós se conectam a nós de grau diferentes dos seus).
Centralização de Grau	Métrica derivada da centralidade de grau que mede o quão central o nó mais central é em relação a todos os outros nós da rede, baseada no grau dos nós.
Centralização de Proximidade	Métrica derivada da centralidade de proximidade (*closeness*) que mede o quão central o nó mais central é em relação a todos os outros nós da rede, baseada na distância existente entre todos os pares de nós.
Diâmetro	Diâmetro da rede (grafo) atual.
Tamanho da Clique Máxima	Tamanho da maior clique do grafo atual, isto é, tamanho do subconjunto máximo de nós no qual todos os elementos do conjunto estão ligados uns aos outros.
Média dos Caminhos Mínimos	Média dos caminhos mínimos entre todos os pares de nós no componente gigante.

Fonte: Digiampietri et al. (2014a)

5.2.2 Análise dos resultados

Inicialmente são apresentadas as redes sociais. A figura 19 apresenta duas nas quais cada nós representa uma cidade que possui ao menos um doutor trabalhando nela. Cada aresta destas redes representa a ligação entre duas cidades (indicando que há doutores trabalhando nestas duas cidades, cujos currículos possuem ao menos uma ligação explícita). Os nós são coloridos de acordo com o estado em que a cidade está situada. A diferença entre as redes é que na da esquerda são apresentadas apenas as arestas entre cidades de um mesmo estado (arestas coloridas). Já na rede da direita, são apresentadas adicionalmente as arestas entre cidades de diferentes estados (arestas cinza). A partir desta figura é possível se ter uma noção inicial da quantidade de cidades que possuem doutores atuando em cada estado brasileiro e da densidade das relações entre cidades.

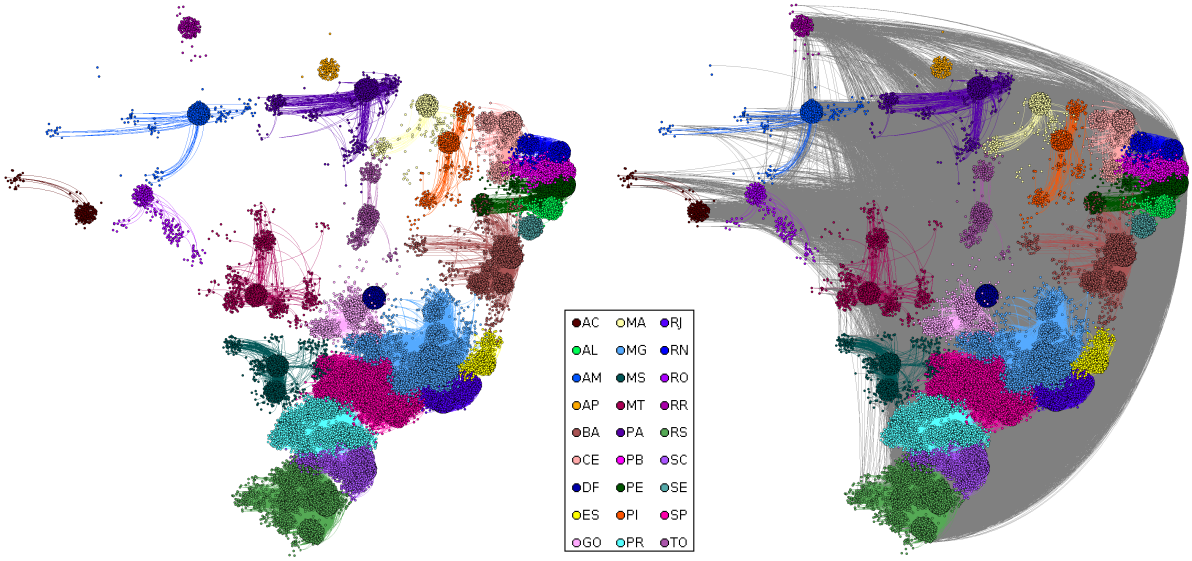
Figura 19 – Rede social dos doutores - cidades



Fonte: Digiampietri et al. (2014a)

Enquanto nas redes da figura 19 cada nó corresponde a uma cidade, nas redes da figura 20, cada nó representa um dos 156.278 doutores da amostra. Estes doutores estão posicionados nas proximidades de sua cidade de atuação. Observa-se a grande quantidade de doutores em algumas regiões do país, principalmente nas capitais e outros grandes centros urbanos. Uma quantificação e detalhamento sobre algumas informações das redes serão apresentados adiante.

Figura 20 – Rede social dos doutores que atuam no Brasil



Fonte: Digiampietri et al. (2014a)

A tabela 14 contém algumas métricas das 28 redes analisadas (nacional, estaduais e do Distrito Federal). Observa-se pelo número de nós (quantidade de doutores) que mais de metade dos doutores estão em São Paulo (30,3%), Rio de Janeiro (13,2%) e Minas Gerais (9,5%). Adicionando-se os doutores do Rio Grande do Sul, Paraná, Pernambuco e Santa Catarina obtêm-se quase 75% de todos os doutores da rede. Por outro lado, ao se somarem os doutores do Amapá, Roraima, Acre, Rondônia e Tocantins obtêm-se menos 1% do total (DIGIAMPIETRI et al., 2014a).

Tabela 14 – Métricas calculadas para cada rede social produzida

	Nós	Arestas	Nós no Componente Gigante	Porcentagem de Nós no Componente Gigante	Densidade	Grau Médio	Coeficiente de Clusterização	Assortatividade de Grau	Centralidade de Grau	Centralidade de Proximidade	Diâmetro	Tamanho da Clique Máxima	Média dos Caminhos Mínimos
AC	317	110	24	7,57%	0,00220	0,694	0,342	0,164	0,071	0,003	8	4	3,19
AL	1109	850	379	34,17%	0,00138	1,533	0,214	-0,053	0,096	0,003	13	5	5,40
AM	1583	1595	706	44,60%	0,00127	2,015	0,195	-0,052	0,043	0,002	13	7	5,17
AP	161	55	40	24,84%	0,00427	0,683	0,057	-0,308	0,424	0,039	7	3	2,85
BA	5357	7291	2570	47,97%	0,00051	2,722	0,182	0,021	0,027	0,001	19	8	5,72
CE	3561	8288	2195	61,64%	0,00131	4,655	0,209	0,076	0,039	0,001	19	10	5,14
DF	5421	6791	2567	47,35%	0,00046	2,505	0,191	0,026	0,020	0,001	19	8	5,68
ES	1969	1832	843	42,81%	0,00095	1,861	0,251	0,033	0,040	0,002	19	7	6,91
GO	2987	4071	1458	48,81%	0,00091	2,726	0,232	0,039	0,029	0,001	20	11	5,69
MA	1154	807	334	28,94%	0,00121	1,399	0,323	0,208	0,043	0,003	17	6	5,83
MG	15234	45208	10137	66,54%	0,00039	5,935	0,158	0,067	0,021	0,000	21	11	5,31
MS	1828	2058	808	44,20%	0,00123	2,252	0,188	0,002	0,044	0,002	16	6	4,95
MT	1718	1182	585	34,05%	0,00080	1,376	0,217	0,062	0,044	0,002	17	7	6,05
PA	2576	3399	1405	54,54%	0,00102	2,639	0,194	-0,104	0,054	0,001	15	8	5,68
PB	3488	5163	1741	49,91%	0,00085	2,960	0,208	-0,031	0,042	0,001	17	9	5,32
PE	4842	10060	3106	64,15%	0,00086	4,155	0,178	0,036	0,028	0,001	18	8	5,13
PI	1026	774	345	33,63%	0,00147	1,509	0,212	-0,084	0,115	0,004	12	6	4,68
PR	10307	21472	6577	63,81%	0,00040	4,166	0,185	0,013	0,014	0,000	16	11	5,75
RJ	20639	50368	13638	66,08%	0,00024	4,881	0,146	0,081	0,012	0,000	18	18	5,62
RN	2627	3413	1372	52,23%	0,00099	2,598	0,213	-0,072	0,036	0,001	16	8	5,75
RO	411	203	92	22,38%	0,00241	0,988	0,320	-0,015	0,090	0,008	10	5	4,67
RR	279	62	15	5,38%	0,00160	0,444	0,287	-0,077	0,128	0,004	5	4	2,22
RS	13012	39254	9621	73,94%	0,00046	6,034	0,167	0,080	0,019	0,000	16	13	5,28
SC	5578	9751	3376	60,52%	0,00063	3,496	0,181	0,030	0,022	0,001	22	8	5,60
SE	1193	1165	485	40,65%	0,00164	1,953	0,255	0,117	0,058	0,003	15	8	5,45
SP	47314	179092	35836	75,74%	0,00016	7,570	0,130	0,057	0,007	0,000	18	14	5,24
TO	587	267	116	19,76%	0,00155	0,910	0,323	0,150	0,101	0,007	11	7	4,63
Brasil	**156278**	**641825**	**118678**	**75,94%**	**0,00005**	**8,214**	**0,116**	**0,054**	**0,004**	**0,000**	**18**	**24**	**5,50**

Fonte: Digiampietri et al. (2014a)

A *porcentagem de nós no componente gigante* contém a porcentagem dos nós que pertencem ao maior componente conexo de cada rede. Conforme apresentado (seção 2.1), valores altos para esta métrica são considerados positivos, pois indicam que grande parte dos indivíduos está dentro do maior fluxo de conhecimento/informação da rede. Observa-se que quase 76% dos doutores estão na componente gigante da rede nacional. Porém, ao se considerar as redes estaduais, apenas nas redes de São Paulo e do Rio Grande do Sul mais de 70% dos doutores pertencem ao componente gigante. Esta diferença indica que ligações

entre estados são bastante importantes para a conectividade da rede nacional e, mesmo, para conectar indiretamente doutores de um mesmo estado.

A *densidade* de todas as redes é bastante baixa, fato que é comum em redes sociais acadêmicas que envolvem grandes números de pessoas. O *grau médio* da rede nacional é de 8,2. As redes estaduais com maior grau médio são as de São Paulo, Rio Grande do Sul e Minas Gerais.

Conforme apresentado, o *coeficiente de clusterização* costuma ser associado a estabilidade (ou maturidade) de uma rede (LEMIEUX; OUIMET, 2008). Na rede nacional, esta métrica vale apenas 0,116. A métrica possui valores acima de 0,3 apenas para três redes estaduais (todas relativamente pequenas): Acre, Tocantins, Maranhão e Rondônia.

As métricas que envolvem caminhos/distâncias (como diâmetro) foram calculadas considerando-se apenas o componente gigante.

As medidas de *centralização de grau e de proximidade* (também conhecidas como centralidade da rede) são baseadas nas medidas de centralidade dos nós e servem para indicar o quão importante o nó mais central de cada rede é para a sua rede (com base, respectivamente, em seu grau ou nos caminhos médios mínimos entre os nós). Centralizações muito altas em redes sociais costumam não ser bem vistas, pois indicam que o indivíduo mais central da rede é muito importante (ou influente) para esta rede (ou seja, a rede depende muito dele). Os menores valores de centralização são encontrados em São Paulo, Rio de Janeiro e Paraná. Já a rede do Amapá possui os maiores valores para estas métricas.

O *diâmetro* da rede nacional é igual a 18, o que pode indicar que a transmissão de conhecimento/informação na rede pode ser um pouco lenta. Os menores valores de diâmetro nas redes estaduais encontram-se todos em redes pequenas: Roraima (5), Amapá (7) e Acre (8). Já os maiores valores para o diâmetro ocorrem nas redes de Santa Catarina (22), Minas Gerais (21) e Goiás (20). Destaca-se que o diâmetro da rede nacional é menor do que o diâmetro de sete redes estaduais, indicando novamente que há ligações entre estados que encurtam os caminhos de algumas redes nacionais. A *Média dos Caminhos Mínimos* tem seu menor valor (assim como ocorreu com o diâmetro) nas redes de Roraima, Amapá e Acre. Os maiores valores ocorrem nas redes do Espírito Santo, Mato Grosso e Maranhão. Na rede nacional, este valor é de 5,5.

O tamanho do maior *clique* em cada rede pode indicar um conjunto coeso de doutores. Na rede nacional há um clique com tamanho 24 (a qual contém docentes de

diferentes estados). Os maiores cliques estaduais estão nas redes do Rio de Janeiro (18 doutores), São Paulo (14) e Rio Grande do Sul (13).

A fim de analisar o relacionamento dos doutores entre diferentes estados, foi estudada a natureza geográfica de cada ligação entre doutores de forma a se verificar se envolve doutores de um mesmo estado ou não. A tabela 15 apresenta a porcentagem de arestas existentes entre cada par de estados (DIGIAMPIETRI et al., 2014a). As linhas e colunas contêm estados e os valores nas células correspondem a porcentagem de arestas que ligam o estado representado pela respectiva linha com o estado representado pela coluna. Esta porcentagem é calculada em função do número total de arestas do estado representado pela linha desta célula. Assim, a soma de porcentagens de cada linha totaliza 100%. A cor de fundo das células tem sua intensidade proporcional ao valor da célula. Nas últimas três colunas da tabela são apresentados, respectivamente, o total de arestas envolvendo ao menos um doutor do respectivo estado, a porcentagem destas arestas que ocorrem dentro no mesmo estado e a porcentagem de arestas que ligam um doutor do respectivo estado com um doutor de outro estado.

Tabela 15 – Porcentagem de arestas de acordo com o estado

	AC	AL	AM	AP	BA	CE	DF	ES	GO	MA	MG	MS	MT	PA	PB	PE	PI	PR	RJ	RN	RO	RR	RS	SC	SE	SP	TO	Total	Intra	Inter
AC	9%	1%	2%	0%	3%	3%	4%	1%	1%	0%	18%	1%	2%	2%	2%	1%	1%	5%	6%	2%	1%	0%	4%	2%	1%	28%	0%	1196	9%	91%
AL	0%	17%	1%	0%	3%	4%	2%	1%	1%	1%	7%	0%	1%	1%	7%	11%	0%	3%	9%	2%	0%	0%	4%	2%	3%	22%	0%	5033	17%	83%
AM	0%	0%	22%	0%	2%	2%	3%	1%	1%	1%	8%	1%	2%	4%	2%	2%	0%	5%	8%	1%	1%	0%	3%	2%	1%	27%	0%	7298	22%	78%
AP	0%	1%	1%	10%	1%	1%	2%	1%	2%	1%	9%	1%	0%	14%	3%	3%	0%	7%	7%	2%	1%	1%	3%	1%	1%	28%	0%	575	10%	90%
BA	0%	1%	0%	0%	29%	2%	3%	1%	1%	0%	11%	1%	1%	1%	2%	4%	1%	3%	8%	2%	0%	0%	4%	2%	1%	23%	0%	25155	29%	71%
CE	0%	1%	1%	0%	2%	38%	2%	0%	1%	1%	5%	0%	1%	2%	4%	4%	2%	2%	6%	4%	0%	0%	3%	1%	1%	17%	0%	21582	38%	62%
DF	0%	0%	1%	0%	3%	2%	27%	1%	5%	0%	11%	1%	1%	1%	1%	2%	0%	4%	7%	1%	0%	0%	5%	2%	1%	23%	0%	25375	27%	73%
ES	0%	0%	1%	0%	2%	1%	2%	20%	1%	0%	23%	1%	1%	1%	1%	2%	0%	3%	15%	1%	0%	0%	3%	1%	0%	21%	0%	9392	20%	80%
GO	0%	0%	0%	0%	1%	1%	8%	1%	25%	0%	15%	1%	1%	1%	1%	1%	0%	4%	5%	1%	0%	0%	3%	1%	0%	27%	1%	16211	15%	85%
MA	0%	1%	1%	0%	3%	6%	2%	1%	1%	17%	8%	1%	0%	3%	5%	3%	2%	4%	6%	3%	0%	0%	2%	1%	1%	30%	0%	4667	17%	83%
MG	0%	0%	1%	0%	3%	1%	3%	2%	2%	0%	44%	1%	1%	1%	1%	1%	0%	3%	7%	1%	0%	0%	3%	2%	1%	20%	1%	102228	44%	56%
MS	0%	0%	1%	0%	2%	1%	3%	1%	2%	0%	10%	20%	2%	1%	0%	1%	0%	8%	5%	1%	0%	0%	6%	2%	0%	33%	0%	10169	20%	80%
MT	0%	0%	1%	0%	2%	2%	3%	1%	2%	0%	16%	2%	14%	1%	1%	1%	0%	6%	6%	1%	0%	0%	7%	2%	1%	30%	0%	8467	14%	86%
PA	0%	1%	2%	1%	2%	3%	3%	1%	1%	1%	8%	1%	1%	29%	2%	2%	1%	3%	8%	1%	0%	0%	4%	2%	1%	23%	0%	11655	29%	71%
PB	0%	2%	1%	0%	3%	5%	2%	1%	1%	1%	6%	0%	1%	1%	29%	12%	2%	2%	4%	7%	0%	0%	3%	1%	2%	14%	0%	17845	29%	71%
PE	0%	2%	1%	0%	4%	3%	2%	1%	1%	0%	5%	0%	0%	1%	8%	37%	1%	2%	6%	3%	0%	0%	3%	1%	2%	16%	0%	26932	37%	63%
PI	0%	0%	1%	0%	3%	11%	2%	0%	1%	2%	10%	1%	1%	1%	6%	7%	16%	2%	4%	3%	0%	0%	2%	1%	1%	24%	1%	4905	16%	84%
PR	0%	0%	1%	0%	1%	1%	2%	0%	1%	0%	5%	1%	1%	1%	1%	1%	0%	37%	4%	1%	0%	0%	6%	6%	1%	28%	0%	57474	37%	63%
RJ	0%	0%	1%	0%	2%	1%	2%	1%	1%	0%	8%	1%	1%	1%	1%	2%	0%	3%	53%	1%	0%	0%	4%	2%	0%	15%	0%	95558	53%	47%
RN	0%	1%	1%	0%	3%	6%	2%	0%	1%	1%	6%	0%	0%	1%	8%	6%	1%	3%	6%	25%	0%	0%	4%	2%	1%	21%	0%	13861	25%	75%
RO	1%	0%	2%	0%	2%	3%	5%	1%	2%	0%	13%	1%	2%	1%	2%	1%	0%	6%	8%	1%	10%	0%	7%	2%	0%	28%	0%	1977	10%	90%
RR	0%	0%	3%	0%	1%	2%	3%	1%	1%	1%	23%	1%	1%	2%	5%	3%	1%	4%	9%	3%	1%	6%	6%	2%	1%	20%	0%	1067	6%	94%
RS	0%	0%	0%	0%	1%	1%	2%	0%	1%	0%	4%	1%	1%	1%	1%	1%	0%	5%	5%	1%	0%	0%	53%	8%	1%	14%	0%	73756	53%	47%
SC	0%	0%	1%	0%	1%	1%	2%	0%	1%	0%	5%	1%	1%	1%	1%	1%	0%	11%	6%	1%	0%	0%	18%	31%	1%	17%	0%	31159	31%	69%
SE	0%	2%	1%	0%	5%	3%	2%	1%	1%	1%	10%	1%	1%	1%	4%	6%	1%	4%	6%	2%	0%	0%	5%	2%	16%	26%	0%	7448	16%	84%
SP	0%	0%	1%	0%	2%	1%	2%	1%	1%	0%	7%	1%	1%	1%	1%	1%	0%	5%	5%	1%	0%	0%	3%	2%	1%	61%	0%	295446	61%	39%
TO	0%	0%	1%	0%	2%	3%	3%	1%	4%	0%	20%	1%	1%	1%	3%	2%	1%	6%	5%	1%	0%	0%	6%	3%	1%	24%	10%	2638	10%	90%

Fonte: Digiampietri et al. (2014a)

Duas informações têm maior destaque nesta tabela: as células mais escuras na diagonal principal (indicando arestas que ligam dois doutores de um mesmo estado) e algumas colunas com células predominantemente mais escuras (que indicam que alguns estados atraem muitas ligações para eles, em especial os estados que possuem as maiores redes de doutores, como o caso de São Paulo, Minhas Gerais, Rio de Janeiro e Rio Grande do Sul). A primeira informação indica uma assortatividade de estado (NEWMAN,

2003a), isto é, há uma predominância entre arestas de doutores de um mesmo estado. Excetuando-se estas informações iniciais, observa-se que muitas das demais células mais escuras (porcentagens maiores) ocorrem entre estados geograficamente próximos, por exemplo, 14% das arestas de doutores do Amapá os ligam a doutores do Pará, 12% das arestas da Paraíba são com doutores de Pernambuco e 11% das arestas do Piauí são com o Ceará.

Observa-se pelas últimas colunas da tabela 15 que apenas 9% das arestas do Acre envolvem dois doutores deste estado. No outro extremo, mais de 60% das arestas de São Paulo ocorrem entre dois doutores de São Paulo.

Além da grande diversidade de tamanho e características das redes de doutores dos estados brasileiros, existem também diferenças significativas quanto às áreas de atuação destes doutores. Para se analisar este fato, optou-se por investigar apenas os doutores que cadastraram uma única grande-área de atuação em seus Currículos Lattes. Dos 156.278 doutores da amostra, 103.378 satisfizeram este critério.

A tabela 16 apresenta a porcentagem destes doutores em cada estado de acordo com a grande área de atuação. A soma das porcentagens de cada linha atinge 100%. Em cada coluna é destacada a maior porcentagem do respectivo estado. É interessante observar que apesar da área Ciências da Saúde ser a que possui maior número de doutores nesta amostra (quase 20%), apenas no estado de São Paulo ela é a área que, relativamente, apresenta mais doutores (mais de 41% dos doutores do estado). No extremo oposto, Ciências Agrárias é a área que se destaca numa maior quantidade de estados. Ciências Biológicas tem destaque no Rio de Janeiro, Pará e Amazonas e Engenharias no Espírito Santo e Santa Catarina, apenas para citar alguns destaques.

O componente gigante da rede social composta apenas pelos doutores que registraram uma única grande área de atuação pode ser visualizado na figura 21. Assim como nas demais redes deste tipo apresentadas no presente documento, todos os nós tentam se afastar uns dos outros, porém aqueles que possuem uma aresta têm uma força de atração. É possível observar agrupamentos de nós (doutores) de uma mesma cor (isto é, que atuam numa mesma área). Além disso, é interessante observar as bordas destes agrupamentos assim como as regiões nas quais duas áreas se misturam. Destaca-se uma mistura entre doutores atuando em Ciências Biológicas e Ciências Agrárias, e entre Ciências Exatas e da Terra e Engenharias.

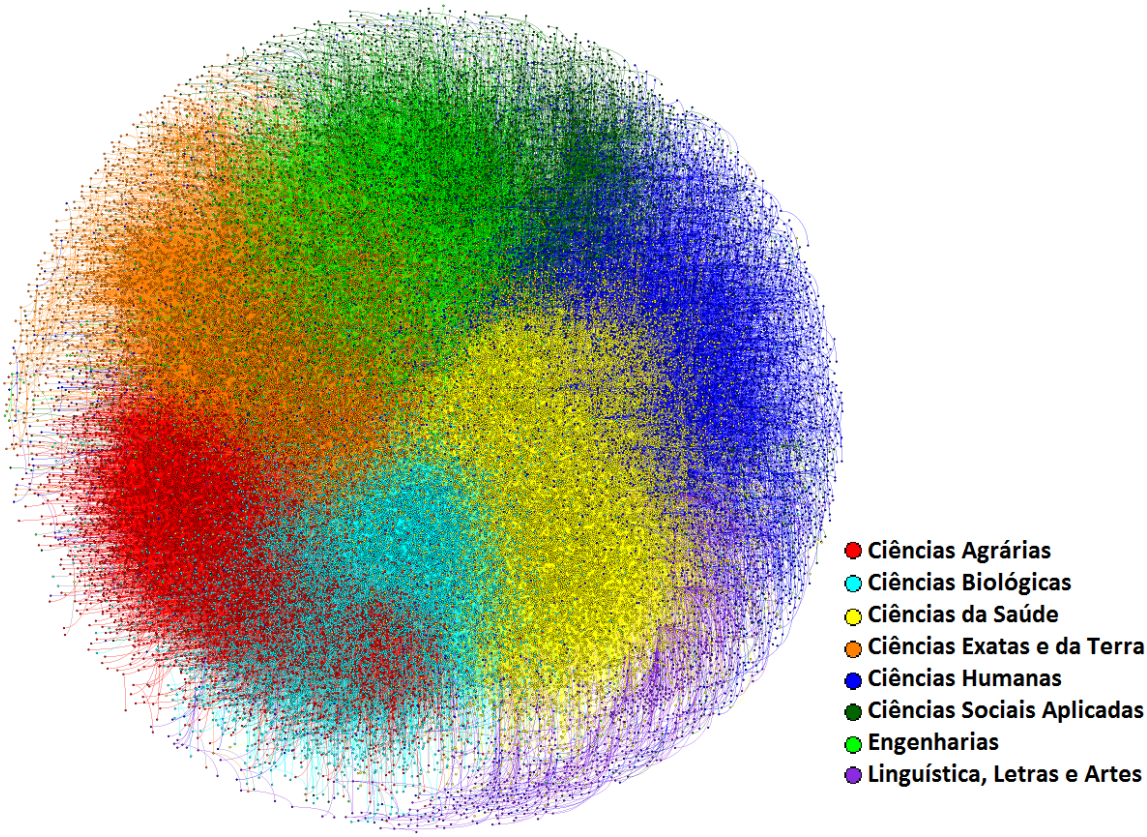
Tabela 16 – Distribuição das áreas de atuação dos doutores pelos estados

	Porcentagem do Total	Ciências Agrárias	Ciências Biológicas	Ciências da Saúde	Ciências Exatas e da Terra	Ciências Humanas	Ciências Sociais Aplicadas	Engenharias	Linguística, Letras e Artes
AC	0,22%	**0,56%**	0,24%	0,12%	0,14%	0,32%	0,10%	0,04%	0,34%
AL	0,72%	0,83%	0,53%	0,60%	0,95%	0,70%	0,78%	0,47%	**1,03%**
AM	1,03%	1,17%	**2,13%**	0,80%	1,21%	0,84%	0,57%	0,73%	0,46%
AP	0,08%	0,11%	0,09%	0,06%	0,11%	**0,12%**	0,09%	0,01%	0,05%
BA	3,49%	3,63%	3,28%	3,33%	3,61%	4,09%	2,82%	2,12%	**5,75%**
CE	2,29%	**2,76%**	1,43%	2,44%	2,45%	2,61%	1,88%	2,38%	2,01%
DF	3,23%	2,87%	3,09%	2,21%	2,83%	4,03%	**5,84%**	2,61%	3,13%
ES	1,25%	1,50%	1,15%	0,95%	1,27%	1,24%	1,34%	**1,69%**	1,19%
GO	2,00%	**3,50%**	1,82%	1,55%	1,85%	2,70%	1,09%	1,35%	2,20%
MA	0,75%	**1,04%**	0,57%	0,74%	0,80%	1,00%	0,56%	0,52%	0,46%
MG	9,88%	**13,74%**	9,67%	8,30%	9,44%	8,67%	9,61%	11,56%	11,01%
MS	1,23%	**2,40%**	1,31%	0,84%	1,08%	1,67%	0,77%	0,46%	1,58%
MT	1,13%	**2,74%**	1,03%	0,70%	0,98%	1,39%	0,65%	0,46%	1,46%
PA	1,64%	2,11%	**2,26%**	0,91%	1,80%	1,92%	1,14%	1,53%	1,68%
PB	2,29%	2,92%	1,26%	1,64%	2,23%	2,84%	1,85%	3,28%	**3,32%**
PE	3,10%	**3,82%**	3,34%	3,25%	3,11%	2,85%	2,95%	2,80%	2,13%
PI	0,70%	**1,49%**	0,61%	0,63%	0,68%	0,85%	0,30%	0,27%	0,81%
PR	6,77%	**9,13%**	6,09%	5,71%	6,19%	7,26%	7,72%	6,16%	7,09%
RJ	12,65%	4,95%	**16,20%**	9,69%	16,09%	13,25%	12,53%	15,20%	14,19%
RN	1,67%	1,28%	1,30%	1,27%	**2,27%**	1,89%	1,39%	2,19%	1,99%
RO	0,24%	**0,40%**	0,30%	0,11%	0,21%	0,35%	0,22%	0,08%	0,36%
RR	0,17%	**0,45%**	0,14%	0,01%	0,17%	0,21%	0,17%	0,09%	0,24%
RS	8,40%	9,80%	6,52%	8,72%	7,31%	9,09%	**10,25%**	6,78%	9,40%
SC	3,42%	3,29%	2,11%	2,85%	2,87%	3,83%	4,32%	**5,69%**	3,37%
SE	0,76%	0,86%	0,52%	0,68%	0,92%	**0,98%**	0,66%	0,62%	0,64%
SP	30,55%	21,69%	32,69%	**41,79%**	29,19%	24,83%	30,04%	30,73%	23,65%
TO	0,36%	**0,95%**	0,31%	0,12%	0,23%	0,48%	0,36%	0,18%	0,46%
# de Doutores	103.378	10.895	13.162	20.516	15.695	16.986	10.196	10.106	5.822

Fonte: Digiampietri et al. (2014a)

Para se ter uma ideia inicial acerca dos conteúdos publicados pelos doutores, foram utilizadas técnicas de mineração de textos para identificar quais as expressões mais utilizadas nos títulos das publicações em cada estado. Foram identificados a partir dos currículos dos doutores um total de 4.566.023 artigos completos (publicados em periódicos ou anais de eventos). Os títulos dos artigos foram filtrados, passando-se todas as letras para minúsculo, removendo-se os acentos e *stop-words*. Em seguida, foram calculadas a frequência das palavras e expressões (DIGIAMPIETRI et al., 2014a).

Figura 21 – Rede de social dos doutores - nós coloridos de acordo com a grande-área de atuação



Fonte: Digiampietri et al. (2014a)

As figuras 22, 23 e 24 contêm as nuvens de palavras e expressões de duas e três palavras considerando todos os títulos analisados[2].

Pelo fato das nuvens de palavras e expressões envolverem um grande número de publicações de diferentes áreas, observa-se a presença de palavras ou expressões gerais utilizadas nos títulos de diferentes áreas (como “estudo”, “análise”, “características”, “estudo caso”) incluindo nomes de cidades ou estados (“rio janeiro”, “sao paulo”, “minas gerais”), mas também é possível encontrar algumas expressões mais específicas (como “redes neurais artificiais” e “soja glycine max”).

Para contextualizar estas palavras e expressões nos estados, foi realizada uma análise da frequência relativa delas em cada estado em relação a suas respectivas frequências em todo o país. A tabela 17 apresenta as cinco expressões de duas palavras relativamente mais frequentes nos títulos dos artigos de cada estado. A grande maioria das células da tabela está preenchida com expressões relacionada ao estado (nome do estado, de alguma

[2] As nuvens foram geradas utilizando a ferramenta Word Cloud Generator (https://www.jasondavies.com/wordcloud/).

Figura 22 – Nuvem de palavras dos títulos das publicações nacionais

Fonte: adaptado de Digiampietri et al. (2014a)

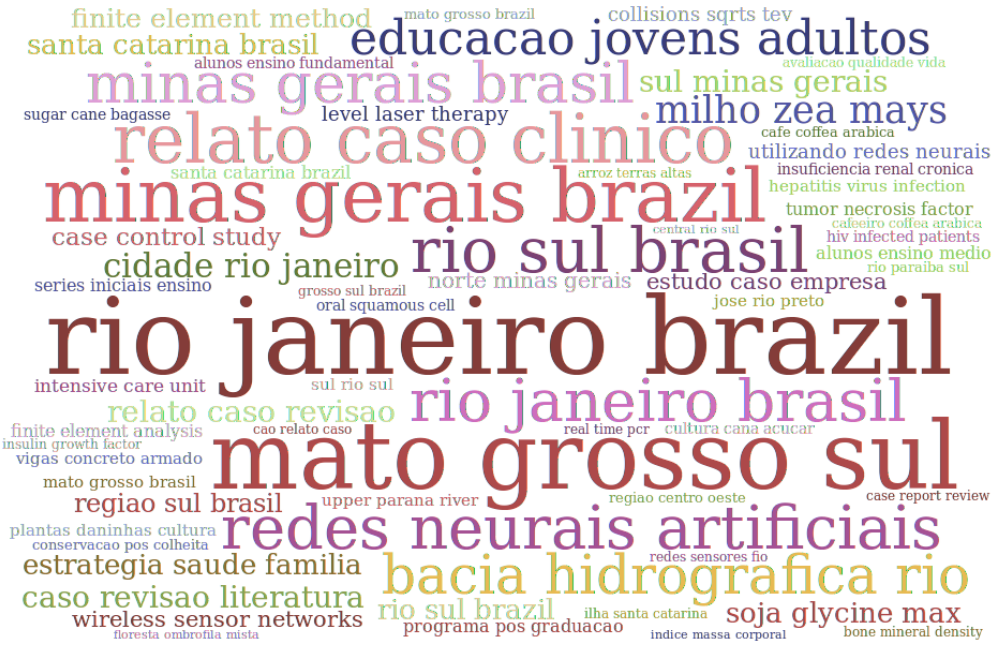
Figura 23 – Nuvem de expressões de duas palavras dos títulos das publicações nacionais

Fonte: adaptado de Digiampietri et al. (2014a)

cidade importante, de um rio do estado ou de uma instituição). Mesmo assim é possível identificar alguns assuntos de interesse: *inflammatory analgesic* no Amapá, *spotted wilt* no Distrito Federal, *cyprinodontiformes rivulidae* no Rio de Janeiro, e *forragem composicao* em Roraima[3] (DIGIAMPIETRI et al., 2014a).

[3] Só foram consideradas expressões que ocorressem ao menos cem vezes no conjunto total de títulos.

Figura 24 – Nuvem de expressões de três palavras títulos das publicações nacionais



Fonte: adaptado de Digiampietri et al. (2014a)

Tabela 17 – Expressões relativamente mais frequentes em cada estado

AC	acre brasil	branco acre	rio branco	acre brazil	chico mendes
AL	cidade maceio	maceio alagoas	mundau manguaba	alagoas brasil	alagoas brazil
AM	industrial manaus	cidade manaus	municipio manaus	manaus amazonas	amazonia central
AP	amapa brasil	amapa brazil	inflammatory analgesic	peso comprimento	osteichthyes characidae
BA	federal bahia	vitoria conquista	feira santana	santana bahia	reconcavo bahia
CE	federal ceara	cidade fortaleza	metropolitana fortaleza	municipio fortaleza	juazeiro norte
DF	universidade brasilia	distrito federal	brasilia brazil	federal district	spotted wilt
ES	hymenoptera bethylidae	cidade vitoria	sul espirito	vitoria espirito	municipio vitoria
GO	municipio jatai	municipio goiania	cidade goiania	federal goias	goiania goias
MA	cidade luis	ilha luis	luis maranhao	maranhao brasil	municipio luis
MG	clinicas ufmg	puc minas	municipio uberlandia	montes claros	federal minas
MS	kaiowa guarani	regiao dourados	sul mato	campo mato	ponta pora
MT	municipio cuiaba	tangara serra	cuiaba mato	chapada guimaraes	rio cuiaba
PA	metropolitana belem	sudeste paraense	nordeste paraense	cidade belem	municipio belem
PB	cidade campina	federal paraiba	sertao paraibano	joao pessoa	municipio campina
PE	cidade recife	cabo santo	jaboatao guararapes	serra talhada	federal pernambuco
PI	teresina piaui	piaui brasil	serra capivara	piaui brazil	cutias dasyprocta
PR	originado projeto	litoral parana	francisco beltrao	campo mourao	federal parana
RJ	250 gev	cyprinodontiformes rivulidae	puc rio	noroeste fluminense	favelas rio
RN	cidade natal	natal brasil	municipio mossoro	municipio natal	natal brazil
RO	porto velho	velho rondonia	rondonia brasil	rio madeira	pastagens degradadas
RR	forragem composicao	roraima brasil	pastagens degradadas	roraima brazil	andropogon gayanus
RS	universidade passo	clinicas porto	universitario santa	rio sinos	embrapa clima
SC	municipio florianopolis	florianopolis brasil	ilha santa	planalto catarinense	balneario camboriu
SE	federal sergipe	aracaju sergipe	sergipe brazil	sergipe brasil	hyptis pectinata
SP	clinico radiografica	reserva parque	sentry antimicrobial	clinicas fmusp	liquido cefalorraqueano
TO	porto nacional	sul tocantins	tocantins brasil	vacas descarte	pastagem cultivada

Fonte: adaptado de Digiampietri et al. (2014a)

5.2.3 Conclusões - doutores atuando no Brasil

Esta pesquisa realizou uma análise geral da rede de doutores que atuam no Brasil, destacando alguns aspectos das redes estaduais e da rede nacional.

Observou-se grande heterogeneidade das redes tanto considerando as características estruturais quanto em relação à distribuição de doutores por grandes-áreas, expressões nos títulos das publicações e relacionamentos entre estados. Uma discussão mais detalhada sobre esta pesquisa e alguns pontos de comparação com trabalhos correlatos pode ser encontrada em Digiampietri et al. (2014a).

5.3 Conclusões

Neste capítulo foram realizados dois tipos de análises sobre redes sociais acadêmicas. Na primeira análise, diferentes medidas bibliométricas e da análise de redes sociais foram utilizadas para caracterizar e comparar os programas de pós-graduação em Ciência da Computação no Brasil. Um dos resultados interessantes foi a relação entre essas medidas puramente quantitativas e as notas atribuídas pela avaliação da CAPES dos programas de pós-graduação (atividade extremamente complexa e que envolve análises quantitativas e qualitativas).

A segunda análise realizada foi mais superficial, por envolver currículos de um volume muito maior de pessoas, e focou em medidas da análise de redes sociais e uso de técnicas de mineração de texto para conseguir caracterizar/distinguir as redes acadêmicas de cada um dos estados brasileiros. Constatou-se que a grande diversidade social e cultural brasileira também ocorre nas redes acadêmicas. Também foi possível verificar que as ligações entre doutores de estados diferentes tipicamente ocorrem ou entre estados geograficamente próximos ou entre um estado qualquer e um dos estados que possuem uma das maiores redes acadêmicas.

6 Predição de relacionamentos

Conforme apresentado na seção 2.12, a predição de relacionamentos tem sido bastante estudada nos últimos anos e pode ser aplicada a diferentes áreas.

O problema de predição de relacionamentos pode ser dividido em predição de relacionamentos novos/inéditos (isto é, prever quais pares de pessoas que nunca se relacionaram numa rede social irão começar a se relacionar) e no problema geral de predição de relacionamentos (predizer que pares de pessoas irão se relacionar independentemente delas já terem ou não se relacionado). Tipicamente, a expressão "predição de relacionamentos" (ou *link prediction*) se refere ao primeiro problema.

Na presente pesquisa, o problema de predição de relacionamentos foi aplicado a redes sociais de coautoria, visando a prever colaborações na publicação de coautorias (tanto colaborações inéditas como reincidentes). A predição foi tratada como um problema de classificação em inteligência artificial que, para um dado par de pesquisadores, tem o objetivo de classificar esse par como "possuirão relacionamento" ou "não possuirão relacionamento" (ou, especificamente, "serão coautores" ou "não serão coautores").

Para isto, um conjunto de atributos/características foi extraído dos dados dos pesquisadores, divididos em dois tipos de atributos: os estruturais, extraídos da rede de coautorias; e atributos específicos do domínio extraídos do currículo de cada pesquisador. A solução proposta foi testada utilizando dados de currículos da Plataforma Lattes, porém, é flexível, podendo ser aplicada a diferentes conjuntos de atributos, considerando, por exemplo, apenas atributos estruturais (e neste caso, a única entrada para o algoritmo é uma rede social, representada como uma lista de arestas e um rótulo temporal).

Além do desenvolvimento de uma estratégia para a predição de coautorias, nesta pesquisa um sistema gerenciador de workflows científicos (*Scientific Workflow Management System (SWMS)*) foi utilizado para a construção de um sistema modular de predição de relacionamentos em redes sociais. A base do sistema é a predição considerando a rede social de entrada e, adicionalmente, foram desenvolvidos módulos para a extração de atributos específicos a partir dos dados dos Currículos Lattes.

Os resultados inicias da estratégia de predição de coautorias utilizada e do sistema de predição desenvolvido já foram publicados (DIGIAMPIETRI; SANTIAGO; ALVES, 2013; DIGIAMPIETRI; MARUYAMA, 2014; DIGIAMPIETRI et al., 2015) e serão sumarizados a seguir.

6.1 Metodologia

A pesquisa realizada sobre este tema foi organizada em oito atividades: revisão da literatura correlata; seleção da amostra; cálculo dos atributos; filtragem horizontal dos dados; balanceador do conjunto de treinamento; especificação e desenvolvimento do sistema, execução dos experimentos; e análise dos resultados.

Revisão da literatura correlata. A revisão da literatura foi realizada principalmente para a identificação das técnicas que têm sido utilizadas na predição de relacionamentos em redes sociais e dos atributos (específicos ou estruturais). Um breve resumo sobre a revisão foi apresentado na seção 2.12.

Seleção da amostra. A amostra utilizada neste trabalho corresponde aos dados oriundos dos currículos de 657 docentes permanentes dos programas de pós-graduação em Ciência da Computação com doutorado e/ou mestrado acadêmico que atuaram em ambos triênios: 2004-2006 e 2007-2009. Além do fato de este conjunto de dados já ter sido utilizado em outras pesquisas (ver capítulo 5), esta amostra foi considerada de interesse por conter dados de docentes de uma única área do conhecimento que, potencialmente, possuem diferentes tipos de relacionamentos pertinentes para a predição de coautorias (por exemplo, orientação de alunos em comum, colegas de trabalho, e relações de coautorias).

Cálculo dos atributos. Inicialmente foram calculados atributos específicos de cada par de docentes, sem considerar informações adicionais da rede. Um dos objetivos foi verificar a capacidade preditiva destes atributos (DIGIAMPIETRI; SANTIAGO; ALVES, 2013). Estes atributos foram divididos em períodos de tempo: *passado*, *presente* e *futuro*. A divisão entre passado e presente é feita, pois dados mais atuais tendem a ser mais relevantes para a predição. O sistema é treinado com dados do *passado* e do *presente* visando a predizer os relacionamentos do *futuro*. Para o treinamento, os dados de 1971 a 2000 foram considerados *passados*; de 2001 a 2005, dados atuais (*presente*) e os modelos foram gerados para predizer as coautorias que ocorreram de 2006 a 2010 (*futuro*). Para os testes, todas as janelas de tempo foram deslocadas em cinco anos, assim, o sistema tentou predizer os relacionamentos ocorridos de 2011 a 2015.

A tabela 18 descreve os 19 atributos, sendo os dois primeiros relacionados à classe. *Coautorias a serem preditas* contém o número de coautorias que ocorrerão no período *futuro* (ou seja, aquelas que deseja-se predizer), já o atributo *classe* contém uma indicação se o respectivo par de pessoas colaborará ou não em uma coautoria no futuro. Ou outros

Tabela 18 – Atributos específicos utilizados

Atributo	**Descrição**
coautorias a serem preditas	Quantidade de artigos completos publicados em coautoria pelo par de pesquisadores em análise em conferências ou em periódicos no período *futuro*.
classe	Atributo que assume o valor "possuirão relacionamento" caso o atributo "coautorias a serem preditas" seja maior que zero e, caso contrário, "não possuirão relacionamento".
periódicos anterior	Quantidade de artigos publicados em periódicos em coautoria pelo par de pesquisadores no período *passado*.
conferências anterior	Quantidade de artigos completos publicados em conferências em coautoria pelo par de pesquisadores no período *passado*.
periódicos atual	Quantidade de artigos publicados em periódicos em coautoria pelo par de pesquisadores no período *presente*.
conferências atual	Quantidade de artigos completos publicados em conferências em coautoria pelo par de pesquisadores no período *presente*.
orientação anterior	Atributo que recebe o valor 1 (um) caso um dos pesquisadores tenha sido orientador do outro no período *passado*, ou 0 (zero) caso contrário.
orientação atual	Atributo que recebe o valor 1 (um) caso um dos pesquisadores tenha sido orientador do outro no período *presente*, ou 0 (zero) caso contrário.
orientação em andamento	Atributo que recebe o valor 1 (um) caso um dos pesquisadores seja orientador, em uma orientação em andamento, no período *presente*, ou 0 (zero) caso contrário.
orientadores em comum	Quantidade de orientadores e coorientadores que foram orientadores dos dois pesquisadores em análise.
orientandos em comum	Quantidade de orientandos e coorientandos que foram orientados pelos dois pesquisadores em análise.
vizinhos em comum	Quantidade de vizinhos em comum entre os dois pesquisadores na rede social acadêmica (incluindo os diferentes relacionamentos: coautoria, orientação, etc).
programas em comum	Atributo que recebe o valor 1 (um) caso os dois pesquisadores pertençam ao mesmo programa de pós-graduação, ou 0 (zero) caso contrário.
artigos periódico1	Quantidade de artigos publicados em periódicos no período *presente* pela pessoa 1.
artigos anais1	Quantidade de artigos completos publicados em anais de conferências no período *presente* pela pessoa 1.
artigos periódico2	Quantidade de artigos publicados em periódicos no período *presente* pela pessoa 2.
artigos anais2	Quantidade de artigos completos publicados em anais de conferências no período *presente* pela pessoa 2.
distância	Distância geográfica entre os endereços profissionais do dois pesquisadores.
subáreas em comum	Número de subáreas de atuação que os dois pesquisadores possuem em comum.

Fonte: Digiampietri et al. (2015)

atributos incluem informações sobre artigos completos publicados em conjunto pelo par (em periódicos e em conferências); relação de orientação entre os docentes; existência de orientadores em comum; existência de orientandos em comum; se os dois pertencem ao mesmo programa de pós-graduação; áreas de atuação em comum; distância geográfica de seus endereços profissionais; e quantos vizinhos em comum eles possuem (considerando os *links* explícitos de seus currículos). Destaca-se que este último atributo pode ser considerado estrutural, mas por ter sido obtido diretamente da análise de um par de currículos e também

por considerar outras relações além das de coautoria, foi colocado juntamente com os atributos específicos. Um atributo equivalente, mas calculado considerando-se apenas as informações da rede de coautorias, será apresentado juntamente com os atributos estruturais.

Além dos atributos específicos, foram extraídos 12 atributos estruturais da rede/grafo de coautorias (tabela 19). A escolha destes atributos foi baseada no bom resultado que apresentaram nos trabalhos correlatos (LIBEN-NOWELL; KLEINBERG, 2003; HASAN et al., 2006; MURATA; MORIYASU, 2008; HASAN; ZAKI, 2011; LU; ZHOU, 2011; DHOTE; MISHRA; SHARMA, 2013). Para o atributo estrutural *índice Katz*, foram utilizadas três variações, modificando-se apenas um de seus parâmetros. O atributo estrutural CN (*common neighbors*) foi calculado considerando-se apenas as relações de coautoria do *presente* (de 2001 a 2005), pois diversos trabalhos constataram que quanto mais antigas as relações menos significantes elas serão para a predição (TIAN et al., 2010; CUKIERSKI; HAMNER; YANG, 2011).

Filtragem horizontal dos dados. A combinação par a par das pessoas de uma dada rede social pode gerar um número muito grande de pares a serem analisados. Por exemplo, os 657 docentes da amostra combinados dois a dois resultam em 215.496 pares. Destes pares, apenas uma pequena parcela irá colaborar no futuro (na amostra estudada, menos de 0,5%). Assim, é possível realizar uma filtragem horizontal dos dados (remoção de alguns pares) de forma a se excluir pares que tenham chances muito pequenas de serem coautores. É comum em trabalhos de predição, se excluir pares que não estejam no mesmo componente conexo da rede ou que tenham algumas de suas métricas com valores iguais a zero ou *null* (DIGIAMPIETRI; SANTIAGO; ALVES, 2013). Estes pares excluídos são pré-classificados como "não possuirão relacionamento". Três formas de filtragem de dados foram implementadas. Na primeira, são descartados os pares para os quais todos os valores de atributos são nulos. Na segunda, são eliminados os pares que não estão no mesmo componente conexo. Na terceira forma, o usuário passa como parâmetro uma expressão booleana que será aplicada sobre cada par e indicará se o par deverá ou não ser excluído.

O **balanceamento no conjunto de treinamento** é comumente utilizado em conjuntos de dados desbalanceados e visa a balancear este conjunto, ou seja, equiparar a porcentagem de pares pertencentes à classe "possuirão relacionamento" com aqueles que "possuirão relacionamento". Em redes de coautorias é bastante comum o conjunto de dados estar desbalanceado, mesmo após a filtragem horizontal dos dados. Isto é, haverá muito mais

Tabela 19 – Atributos estruturais utilizados

Atrib.	**Descrição**
classe	Atributo que assume o valor “possuirão relacionamento” ou “não possuirão relacionamento” com base nos dados do intervalo de tempo chamado neste trabalho de *futuro*.
CN	Common Neighbors - número de vizinhos em comum no grafo correspondente à rede social.
SAL	Salton Index - índice que mede a coocorrência de dois elementos dividida pela raiz quadrada da multiplicação da ocorrência de cada elemento. Em redes sociais pode ser usado para medir relação entre o número de vizinhos que duas pessoas têm em comum divido pela raiz quadrada da multiplicação do número de vizinhos de cada um.
JAC	Jaccard’s coefficient - índice que mede a similaridade entre dois conjuntos dividindo o número de elementos da intersecção dos dois conjuntos pelo número de elementos da união (por exemplo, número de vizinhos em comum dividido pela união dos vizinhos de duas pessoas).
AA	Adamic-Adar - índice que atribui peso na relação de duas pessoas favorecendo as relações entre pessoas que possuem poucos relacionamentos (o peso do relacionamento é calculado pela somatória de 1 divido pelo logaritmo do número de relacionamentos [grau] dos vizinhos em comum destas duas pessoas).
RA	Resource Allocation - índice que atribui peso na relação de duas pessoas favorecendo as relações entre pessoas que possuem poucos relacionamentos (o peso do relacionamento é calculado pela somatória de 1 divido pelo número de relacionamentos [grau] dos vizinhos em comum destas duas pessoas).
SOR	Sørensen Index - índice calculado como sendo duas vezes a intersecção entre dois conjuntos dividida pela soma dos elementos de cada conjunto (por exemplo, número de vizinhos em comum dividido pelo número de vizinhos da primeira pessoa mais o número de vizinhos da segunda).
HPI	Hub Promoted Index - índice calculado pela divisão do número de elementos da intersecção de dois conjuntos dividido pelo número mínimo de elementos entre estes dois conjuntos (por exemplo, número de vizinhos em comum de duas pessoas dividido pelo número mínimo de vizinhos destas pessoas).
HDI	Hub Depressed Index - índice calculado pela divisão do número de elementos da intersecção de dois conjuntos dividido pelo número máximo de elementos entre estes dois conjuntos (por exemplo, número de vizinhos em comum de duas pessoas dividido pelo número máximo de vizinhos destas pessoas).
LHM	Leicht-Holme-Newman Index - índice calculado pelo número de elementos da intersecção de dois conjuntos dividido pelo produto do número de elementos de cada conjunto (por exemplo, número de vizinhos em comum dividido pela multiplicação do número de vizinhos de duas pessoas).
PA	Preferential Attachment - índice dado pelo produto entre o número de elementos de dois conjuntos (por exemplo, produto do número de vizinhos de duas pessoas).
KATZ 0,05 0,005 0,0005	Katz é um índice calculado de maneira iterativa para estimar a influência de um par de nós em uma rede considerando-se os caminhos existentes entre os nós. Para este cálculo existe a necessidade da definição de uma constante Beta. Os valores utilizados foram: 0,05 ; 0,005 ; e 0,0005.
SP	Shortest Path - caminho mínimo entre dois nós da rede.

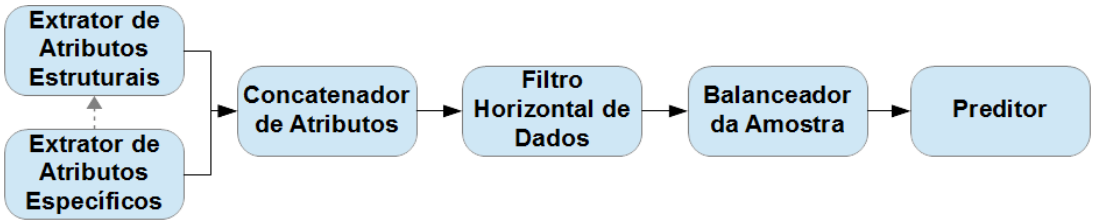
Fonte: Digiampietri et al. (2015)

pares de pessoas cujo valor do atributo classe será “não possuirão relacionamento”. Esse tipo de desbalanceamento, se não tratado, pode afetar significativamente o desempenho do preditor/classificador, afetando principalmente a sensibilidade ou a revocação da solução (HE et al., 2008). Atualmente, apenas uma estratégia de balanceamento foi utilizada: *Random Oversampling*, a qual adiciona aleatoriamente elementos da classe minoritária até que o número de elementos dessa classe se iguale ao número da classe majoritária.

Especificação e desenvolvimento do sistema. O sistema desenvolvido nesta pesquisa utiliza um SWMS, permitindo a fácil incorporação de novos módulos ao sistema e tem as vantagens de diferentes características providas pelo SWMS (como execução paralela ou distribuída das atividades/módulos do workflow). Na literatura correlata é possível encontrar diversos SWMSs (ALTINTAS et al., 2004; TAYLOR et al., 2005; CALLAHAN et al., 2006; GOECKS et al., 2010). Neste trabalho foi utilizada uma das extensões do SWMS chamado WOODSS (MEDEIROS et al., 2005) pelos seguintes motivos: (i) conhecimento prévio do SWMS e de como estendê-lo em diferentes aplicações (DIGIAMPIETRI et al., 2014c; DIGIAMPIETRI et al., 2012); (ii) o SWMS é brasileiro e possui seu código fonte disponível na linguagem de programação Java; (iii) ele permite o uso de diferentes tipos de atividades básicas (módulos) dos workflows (DIGIAMPIETRI et al., 2013). Detalhes sobre a versão atual do SWMS utilizado bem como exemplos de workflows são apresentados na dissertação de mestrado de Santiago (2015).

A figura 25 apresenta a organização dos novos módulos que foram desenvolvidos especificamente para a predição de relacionamentos.

Figura 25 – Módulos do sistema de predição



Fonte: Digiampietri et al. (2015)

O módulo *Extrator de Atributos Estruturais* recebe uma rede social no formato de lista de arestas, em que cada aresta contém um rótulo temporal do ano em que o relacionamento ocorreu e extrai os 12 atributos estruturais descritos na tabela 19.

O módulo *Extrator de Atributos Específicos* recebe como entrada a lista de identificadores de Currículos Lattes e utiliza os arquivos XML dos currículos para calcular os 19 atributos específicos apresentados na tabela 18. Enquanto o primeiro módulo é de propósito geral para a predição de relacionamentos, este é de propósito específico para a predição de coautorias utilizando dados de Currículos Lattes. Adicionalmente, este módulo pode ser utilizado para prover a entrada de dados para o módulo *Extrator de Atributos Estruturais*.

O módulo *Concatenador de Atributos* tabula e concatena os atributos recebidos como entrada. Este módulo também pode ser usado para normalizar os valores de cada atributo para o intervalo de zero a um.

O módulo *Filtro Horizontal de Dados* implementa as três formas de filtragem de dados citadas anteriormente.

Conforme mencionado, o módulo *Balanceador da Amostra* possui atualmente apenas uma estratégia de balanceamento (*Random Oversampling*).

O módulo *Preditor* utiliza um classificador para realizar a predição. Ele tem como entrada: (i) o conjunto de dados a ser utilizado no treinamento e teste, (ii) um parâmetro indicando qual o planejador a ser utilizado, (iii) o caminho (dentro do sistema de arquivos) onde esse planejador se encontra (iv) e a estratégia de validação a ser utilizada. Optou-se por utilizar classificadores disponíveis no ambiente Weka (HALL et al., 2009), que possuem uma interface bem definida. O *Preditor* utiliza reflexão (*reflection*[1]) para invocar o classificador e obter os resultados da classificação.

A maioria dos módulos recebe ainda três intervalos de tempo como parâmetros de entrada, que são chamados de *passado*, *presente* e *futuro*.

Execução dos experimentos. Diferentes experimentos foram realizados de forma a verificar o desempenho da solução proposta, além de testar o desempenho de subconjuntos de atributos tanto para o problema geral de predição de coautorias como para o problema de predição de novas coautorias. Ambas as ações (esta atividade e a análise dos resultados) serão detalhadas na próxima seção.

Análise dos resultados. Os resultados da predição foram analisados de acordo com medidas de precisão e revocação. Além disso, algumas comparações foram realizadas com resultados da literatura. Pelo fato de os dados utilizados neste trabalho não serem os mesmos dos trabalhos correlatos, optou-se por utilizar atributos ou estratégias presentes na literatura e comparar os resultados com a solução proposta. Na próxima seção, são apresentados os resultados utilizando-se apenas um conjunto de atributos específicos ao domínio rede social acadêmica. Na seção seguinte, o resultado do uso de todos os atributos é apresentado e uma comparação com os atributos que se destacaram nos trabalhos correlatos é realizada.

[1] http://docs.oracle.com/javase/tutorial/reflect/

6.2 Experimentos e análise dos resultados

A combinação dois a dois dos 657 pesquisadores utilizados na amostra resulta em 215.496 pares diferentes. Destes, apenas 804 publicaram um ou mais artigos completos em anais ou periódicos no período de 2006 a 2010 (o período considerado como *futuro* nos testes). Este valor corresponde a menos de 0,3731% dos pares e inclui tanto novas relações de coautoria como também a reincidência de relações passadas. Pensando-se apenas na classificação dos pares como "serão coautores" ou "não serão coautores", um classificador que dissesse que todos os pares "não serão coautores", teria mais de 99,6% de chance de acertar. Desta forma, o valor 99,6% será o valor base (*baseline*) para as discussões sobre a classificação como um todo, mas também serão discutidas questões sobre a revocação da predição.

Dois conjuntos de experimentos foram realizados. No primeiro foram utilizados apenas atributos específicos visando a avaliar a capacidade destes atributos de predizerem coautores em um estudo exploratório utilizando diferentes algoritmos de classificação. No segundo, todos os atributos foram considerados e um enfoque maior foi dado à predição de novas coautorias, pois os resultados deste tipo de predição foram insatisfatórios no primeiro conjunto de experimentos, conforme será detalhado a seguir.

6.2.1 Experimentos utilizando apenas atributos específicos

Os experimentos apresentados nesta seção possuem caráter exploratório. Neles, foram utilizados apenas os 13 primeiros atributos específicos da tabela 18 (DIGIAMPIETRI; SANTIAGO; ALVES, 2013).

A primeira atividade realizada foi a aplicação de um filtro horizontal removendo todos os pares de docentes cujos atributos tivessem todos valores iguais a zero (excetuando-se os dois primeiros atributos referentes à classe, pois estes não são avaliados pelo filtro). Com a execução do filtro, restaram 11.800 dos 215.496 pares e, coincidentemente, todos os 804 pares rotulados como "serão coautores" no período de 2006 a 2010 estavam presentes no conjunto remanescente. Os pares excluídos foram classificados como "não serão coautores". Desta forma, para esta amostra, 94,52% dos dados já estariam classificados corretamente e, no restante desta seção será discutida a classificação dos demais 5,48% dos pares.

Dos 11.800 pares selecionados pelo filtro, 804 serão coautores no *futuro* e cerca de 93,19% não serão. Esta porcentagem foi o *baseline* para um teste exploratório sobre os classificadores. Para os testes realizados nesta seção não foi utilizada nenhuma estratégia de balanceamento dos dados de treinamento.

A tabela 20 apresenta os resultados dos dez classificadores que apresentaram melhores resultados no treinamento e teste utilizando-se como parâmetro a porcentagem de acertos da classificação com validação cruzada em 10 subconjuntos. Uma tabela contendo o resultado para 71 classificadores está disponível em Digiampietri, Santiago e Alves (2013). Destes classificadores, 45 obtiveram resultados melhores do que o *baseline*, utilizando-se os valores padrão para seus parâmetros. As taxas de acerto variaram de 94,689% a 60,5493%.

Tabela 20 – Dez melhores resultados dos classificadores testados

Tipo	**N**ome	**T**axa de **a**certo
meta	Bagging	94,689%
meta	EnsembleSelection	94,6526%
meta	RotationForest	94,6162%
trees	FT	94,5526%
meta	Decorate	94,5344%
trees	LMT	94,5071%
trees	J48graft	94,4707%
trees	J48	94,4616%
meta	OrdinalClassClassifier	94,4616%
meta	nestedDichotomies	94,4616%

Fonte: adaptado de (DIGIAMPIETRI; SANTIAGO; ALVES, 2013)

Destaca-se na tabela 20 a presença apenas de metaclassificadores e classificadores baseados em árvores.

A fim de se ter uma visão inicial sobre a utilidade dos atributos extraídos no processo de classificação, duas estratégias foram utilizadas. O uso de seletores de atributos para identificar os subconjuntos de atributos mais importantes na determinação da classe e a análise da correlação entre os valores dos atributos e do atributo classe (neste caso o atributo classe recebeu o valor zero para os pares que não serão coautores e um para os demais).

Os resultados da execução de seletores de atributos[2] são apresentados na tabela 21. Para os seletores que possuem como resultado uma lista ordenada de atributos, o número

[2] Foram utilizadas as implementações dos seletores de atributos disponíveis no arcabouço Weka.

correspondente à posição do atributo na lista aparece na tabela. Para aqueles que apenas selecionam os atributos (sem ordená-los), um "x" aparece na tabela.

Os três atributos que mais se destacam são "conferências atual", "periódicos atual" e "vizinhos em comum" indicando que pessoas que já colaboram tendem a continuar colaborando e pessoas com colaboradores em comum têm maior tendência de colaborarem. Destaca-se que a amostra é composta apenas por docentes credenciados em programas de pós-graduação em Ciência da Computação e estas características da amostra podem ter grande importância nos resultados.

Tabela 21 – Seleção de atributos

	periódicos anterior	conferências anterior	periódicos atual	conferências atual	orientação anterior	orientação atual	orientação em andamento	orientadores em comum	orientandos em comum	vizinhos em comum	programas em comum
FilteredAttributeEval	6	5	3	1	7	9	8	11	4	2	10
ChiSquaredAttributeEval	6	5	3	1	8	9	7	11	4	2	10
GainRatioAttributeEval	4	5	2	1	8	7	3	11	9	6	10
FilteredAttributeEval	6	5	3	1	7	9	8	11	4	2	10
ConsistencySubsetEval	x	x	x	x	x	x	x		x	x	x
CfsSubsetEval			x	x							
SpreadSubsample			x	x							

Fonte: Digiampietri, Santiago e Alves (2013)

A figura 26 contém a correlação entre os valores de todos os atributos. Da primeira linha (correlação entre a classe e os demais atributos) destacam-se: os maiores valores de correlação com os atributos "vizinhos em comum" (0,35), "conferências atual" (0,34) e "periódicos atual" (0,30). É possível observar no restante da tabela que estes três atributos estão altamente correlacionados entre si.

Com base nos resultados dos seletores de atributos e das correlações entre os atributos nove conjuntos de dados utilizando diferentes subconjuntos de atributos foram criados. A tabela 22 contém a indicação de quais atributos foram utilizados em cada um dos conjuntos. Destaca-se que os dois últimos conjuntos foram criados apenas para se verificar a influência da ausência da informação sobre publicações conjuntas anteriores no processo de predição.

A tabela 23 contém os resultados dos dez classificadores selecionados. Os melhores resultados ocorreram com o uso de todos os atributos. Além dos resultados apresentados

Figura 26 – Correlação entre os atributos

Fonte: Digiampietri, Santiago e Alves (2013)

Tabela 22 – Conjuntos de atributos utilizados

	c1	c2	c3	c4	c5	c6	c7	c8	c9
periódicos anterior	x					x	x		
conferências anterior	x					x	x		
periódicos atual	x				x	x	x		
conferências atual	x		x	x	x	x	x		
orientação anterior	x							x	
orientação atual	x							x	
orientação em andamento	x						x	x	
orientadores em comum	x							x	
orientandos em comum	x					x	x	x	x
vizinhos em comum	x	x		x	x	x	x	x	x
programas em comum	x							x	

Fonte: Digiampietri, Santiago e Alves (2013)

na tabela, também foram criados e testados outros onze conjuntos cada um contendo um subconjunto diferente de dez dos onze atributos utilizados. Todos os resultados para estes

subconjuntos foram inferiores ao uso de todos os atributos, indicando assim que todos os atributos contribuem para o processo de predição.

Tabela 23 – Taxa de acerto dos classificadores para cada subconjunto de atributos

	c1	c2	c3	c4	c5	c6	c7	c8	c9
Bagging	94,69%	93,16%	92,63%	93,28%	93,79%	93,78%	93,86%	93,83%	93,45%
EnsembleSelection	94,65%	93,18%	92,72%	93,26%	93,76%	93,79%	93,74%	93,75%	93,41%
RotationForest	94,62%	93,21%	92,63%	93,19%	93,78%	93,82%	93,84%	93,63%	93,28%
FT	94,55%	93,21%	92,63%	93,28%	93,78%	93,89%	93,78%	93,61%	93,47%
Decorate	94,53%	93,19%	92,63%	93,34%	93,71%	93,70%	93,75%	93,55%	93,42%
LMT	94,51%	93,16%	92,63%	93,30%	93,76%	93,82%	93,74%	93,48%	93,44%
J48graft	94,47%	93,21%	92,63%	93,31%	93,70%	93,69%	93,73%	93,60%	93,42%
nestedDichotomies.ND	94,46%	93,21%	92,63%	93,31%	93,70%	93,68%	93,72%	93,62%	93,41%
OrdinalClassClassifier	94,46%	93,21%	92,63%	93,31%	93,70%	93,68%	93,72%	93,62%	93,41%
J48	94,46%	93,21%	92,63%	93,31%	93,70%	93,68%	93,72%	93,62%	93,41%

Fonte: Digiampietri, Santiago e Alves (2013)

Foram realizados alguns testes para verificar a capacidade de alguns algoritmos em predizer a quantidade de publicações em coautoria que um dado par de docentes iria publicar. Para isto, foram utilizadas as informações apenas dos 804 pares que sabidamente irão colaborar. Cada par colaborou, na média, na publicação de 3,57 artigos, sendo a mediana igual a 2 e o desvio padrão igual a 5,79.

Diferentes algoritmos de regressão foram testados de maneira exploratória e utilizando-se os valores-padrão para seus parâmetros. Os resultados destes algoritmos foram analisados de acordo com cinco medidas: coeficiente de correlação, erro absoluto médio, erro quadrático médio, erro absoluto relativo e erro quadrático relativo. A tabela 24 contém o resultado dos algoritmos utilizados[3]. O menor erro absoluto médio relativo foi obtido pelo algoritmo *SMOreg* (67,83%) e o menor valor de erro quadrático relativo pelo algoritmo *Gaussian Processes*.

Tabela 24 – Algoritmos de regressão e resultados

	Gaussian Processes	Isotonic Regression	Least MedSq	Linear Regression	Multilayer Perceptron	Pace Regression	RBF Network	Simple Linear Regression	SMOreg
Correlation coefficient	0,64	0,62	0,20	0,61	0,37	0,63	0,24	0,61	0,64
Mean absolute error	2,34	2,45	2,51	2,42	3,66	2,39	2,93	2,44	2,10
Root mean squared error	4,47	4,54	6,11	4,58	9,50	4,52	5,62	4,60	4,77
Relative absolute error	75,62%	79,03%	80,93%	77,98%	118,18%	77,20%	94,45%	78,67%	67,83%
Root relative squared error	77,00%	78,32%	105,30%	79,00%	163,79%	77,98%	96,93%	79,28%	82,15%

Fonte: Digiampietri, Santiago e Alves (2013)

[3] Foram utilizadas as implementações disponíveis na plataforma Weka

A maioria dos algoritmos de regressão utilizou todos os atributos em sua função. A função mais simples foi criada pelo algoritmo *SimpleLinearRegression* utilizando apenas o atributo "conferências atual": $coautorias\_a\_serem\_preditas = 0,76 * conferencias\_atual + 2,12$ (DIGIAMPIETRI; SANTIAGO; ALVES, 2013).

6.2.2 Experimentos utilizando atributos estruturais e específicos

Nesta subseção são apresentados os resultados que combinaram os atributos estruturais e os específicos. A amostra utilizada é a mesma apresentada anteriormente com o destaque de que, para o cálculo de novas coautorias os pares de docentes que já haviam colaborado em publicações anteriores são descartados.

Para os experimentos desta subseção, foi utilizado o sistema de predição de relacionamentos, cujos módulos foram apresentados na figura 25. A seguir são apresentados os parâmetros utilizados e os resultados obtidos.

Para o treinamento, o intervalo de tempo chamado de *passado* foi de 1971 a 2000; o *presente* de 2001 a 2005 e o *futuro* de 2006 a 2010. O objetivo do sistema foi predizer os relacionamentos ocorridos de 2011 a 2015. Na etapa de filtragem horizontal dos dados foram excluídos todos os pares de elementos para os quais mais de metade dos atributos possuíssem valores nulos. O sistema foi configurado para tratar o problema de predição de relacionamentos novos/inéditos. O classificador selecionado foi o *Rotation Forest*.

Dos 215.496 pares resultantes da combinação dos 657 docentes par a par, apenas 804 serão coautores no *futuro*. A filtragem horizontal dos dados excluiu 193.586 pares (classificados automaticamente como "não serão coautores" e deixando 21.910 pares para a classificação. Destes, 21.074 pares não haviam colaborado em nenhuma publicação até 2010 (ou seja, são candidatos a predição de coautorias novas/inéditas).

A tabela 25 apresenta os resultados da execução sobre o conjunto de dados dos pares que não foram excluídos pelo filtro. O conjunto de treinamento foi balanceado. Além do uso de todos os atributos, são também apresentados os resultados de alguns dos atributos que se destacaram na literatura correlata, bem como do uso de todos os atributos específicos juntos.

Os resultados mostram que a combinação dos atributos utilizados é capaz de predizer com alta taxa de precisão e revocação novas coautorias se destacando sobre o uso de atributos individuais. Porém ainda há muito a se melhorar em relação a revocação da

Tabela 25 – Comparações de Resultados

	Taxa de verdadeiro-positivos	**Taxa de falso-positivos**	**Precisão**	**Revocação**	**F-Measure**	**Área ROC**
Subáreas em comum	0,461	0,358	0,966	0,461	0,614	0,552
Vizinhos em comum	0,486	0,307	0,969	0,486	0,636	0,643
PA	0,742	0,674	0,964	0,742	0,835	0,514
Distância	0,790	0,831	0,961	0,790	0,866	0,540
CN	0,807	0,515	0,969	0,807	0,877	0,646
Katz 0,05	0,823	0,819	0,962	0,823	0,886	0,430
Específicos	0,973	0,963	0,963	0,973	0,968	0,616
Todos	**0,976**	**0,959**	**0,963**	**0,976**	**0,969**	**0,646**

Fonte: adaptado de Digiampietri et al. (2015)

classe positiva.Apesar dos resultados serem inicialmente bons, destaca-se que foi utilizado um conjunto de dados relativamente pequeno (657 pessoas) e homogêneo (formado apenas por professores orientadores permanentes por ao menos dois triênios em programas de pós-graduação em Ciência da Computação). Estas características podem ter simplificado o processo de classificação.

As tabela 26 e 27 comparam o resultado da predição de novas coautorias com o da predição de coautorias (novas e reincidentes). Em ambos os casos foi utilizado o classificador *Rotation Forest*.A precisão, se calculada para todo conjunto de entrada (incluindo os elementos que foram excluídos pelo filtro), é de 99,46% para a previsão de novas coautorias e de 99,65% para a previsão de coautorias.

Tabela 26 – Resultados da predição de novas coautorias

	Taxa de verdadeiro-positivos	**Taxa de falso-positivos**	**Precisão**	**Revocação**	**F-Measure**	**Área ROC**
Não serão coautores	0,995	0,978	0,981	0,995	0,988	0,646
Serão coautores	0,022	0,005	0,080	0,022	0,034	0,646
Média ponderada	0,976	0,959	0,963	0,976	0,969	0,646

Fonte: Digiampietri et al. (2015)

Tabela 27 – Resultados da predição de coautorias (novas e reincidentes)

	Taxa de verdadeiro-positivos	**Taxa de falso-positivos**	**Precisão**	**Revocação**	**F-Measure**	**Área ROC**
Não serão coautores	0,977	0,566	0,972	0,977	0,974	0,823
Serão coautores	0,434	0,023	0,486	0,434	0,459	0,823
Média ponderada	0,951	0,54	0,949	0,951	0,95	0,823

Fonte: Digiampietri et al. (2015)

6.3 Conclusões

Neste capítulo foram apresentados alguns resultados sobre a predição de coautorias (inéditas ou reincidentes).

O problema de predição de coautorias foi tratado como um problema de classificação binária em inteligência artificial que possuía como características para a classificação um conjunto de atributos específicos e gerais relacionados a cada par de pesquisadores que se desejava predizer se iriam ou não colaborar na publicação de um artigo.

Observou-se que a combinação das diferentes características (ou atributos) obteve resultados satisfatórios. Para o conjunto de dados estudado, a precisão da classificação das novas coautorias foi de 97,6% para os dados filtrados e de 99,46% considerando-se o conjunto total de dados. Já para a predição de coautorias (novas e reincidentes) a precisão da classificação foi de 95,1% para o conjunto filtrado de dados e 99,65% para o conjunto total de dados.

Apesar dos resultados promissores, destaca-se que a taxa de revocação para a classe positiva foi baixa (2,2% para a predição de novas coautorias e 43,4% para o problema geral). É necessária a realização de testes e validações adicionais utilizando-se outros conjuntos de dados.

7 Resultados adicionais

Neste capítulo são apresentados outros três tipos de resultados obtidos durante o desenvolvimento desta pesquisa.

Na seção 7.1 são apresentados os resultados referentes à identificação automática das áreas de atuação dos pesquisadores. Na seção 7.2 são apresentados alguns resultados preliminares da combinação de medidas de análise de redes sociais com séries temporais para a predição de tendências. Por fim, na seção 7.3 são apresentados resultados referentes à participação dos orientados na produção de seus orientadores.

7.1 Identificação de áreas de atuação

A identificação automática da área de atuação de um pesquisador ou da área ou assunto de um artigo é uma atividade base para diversas pesquisas que realizam a análise de redes acadêmicas, por exemplo, para avaliação de grupos interdisciplinares, identificação de tendências e busca por especialistas. Esta seção apresenta alguns resultados sobre a combinação de técnicas de mineração de textos (MT) e análise de redes sociais (ARS) na identificação de áreas de atuação (MIYATA; KANO; DIGIAMPIETRI, 2013a).

A ideia principal da pesquisa é avaliar a capacidade de identificação das áreas de atuação (grande-área, área e subárea) dos pesquisadores, com base em suas redes de coautorias e nos títulos de suas publicações. As análises realizadas utilizaram apenas as informações fornecidas pelos possuidores dos currículos, não havendo nenhum tipo de anotação manual dos dados.

7.1.1 Materiais e Métodos

Após uma revisão bibliográfica sobre o assunto, foram selecionadas três medidas a serem testadas e combinadas na identificação das áreas de atuação de pesquisadores utilizando-se dados de Currículos Lattes. Uma delas é baseada na mineração de textos e as outras na análise de redes sociais. Cinco atividades foram realizadas: seleção dos dados; organização das informações de interesse; extração de características; execução dos experimentos; e análise dos resultados.

Seleção dos dados. Para esta parte da pesquisa, optou-se por utilizar apenas os dados dos possuidores de bolsa produtividade em pesquisa. Esta informação foi extraída diretamente dos mais de um milhão de currículos obtidos pela segunda estratégia de identificação (ver seção 3.1). Foram encontrados 13.797 currículos de bolsistas, porém muitos dos pesquisadores cadastram no campo "Áreas de Atuação" mais de uma grande área, área ou subárea. Para a identificação das áreas, optou-se por utilizar apenas os currículos dos pesquisadores que declararam apenas uma grande-área, área ou subárea de atuação, respectivamente. Do total de bolsistas, o número de currículos com uma única grande área é de 9.748 (divididos nas 8 grandes áreas definidas pelo CNPq). 7.297 currículos declararam uma única área de atuação (entre 76 áreas de atuação). Por fim, 3.427 currículos possuíam apenas uma subárea de atuação (entre 443 subáreas diferentes, as quais podiam ser preenchidas manualmente ao invés de retiradas de uma lista como as anteriores). Cada um dos 3 conjuntos foram separados aleatoriamente em 90% dos currículos para treinamento e 10% para testes.

Organização das informações de interesse. As informações dos currículos utilizadas neste trabalho foram: áreas de atuação e informações sobre os artigos publicados em periódicos pelos pesquisadores selecionados no período de 2001 a 2010. As informações das publicações foram utilizadas tanto para montar a rede de coautorias (conforme apresentado na seção 4.1) quanto para a mineração de textos.

Extração de características. Três características foram consideradas para a identificação das áreas de atuação dos pesquisadores. Uma, baseada em **Mineração de Textos**, utilizou a frequência relativa das palavras dos títulos dos artigos (medida TF-IDF - *Term Frequency - Inverse Document Frequency*). Para isto, criou-se um *corpus* com os títulos dos artigos publicados pelos 90% dos pesquisadores utilizados como treinamento. Inicialmente, cada título foi pré-processado com remoção de *stop-words* seguida da execução de um *stemmer* (LOVINS, 1968).

As duas características baseadas em **Análise de Redes Sociais** foram: a porcentagem dos vizinhos pertencentes a cada área (vizinhança nível um) e a porcentagem de vizinhos e vizinhos dos vizinhos (vizinhança nível 2).

Execução dos experimentos. A fim de se avaliar a capacidade das características e de suas combinação para a identificação das áreas de atuação, foram executados experimentos utilizando-se diferentes períodos para os dados de treinamento e de teste (de um a dez anos).

Análise dos resultados. Os resultados foram analisados utilizando-se a medida taxa de acerto e serão apresentados na próxima seção.

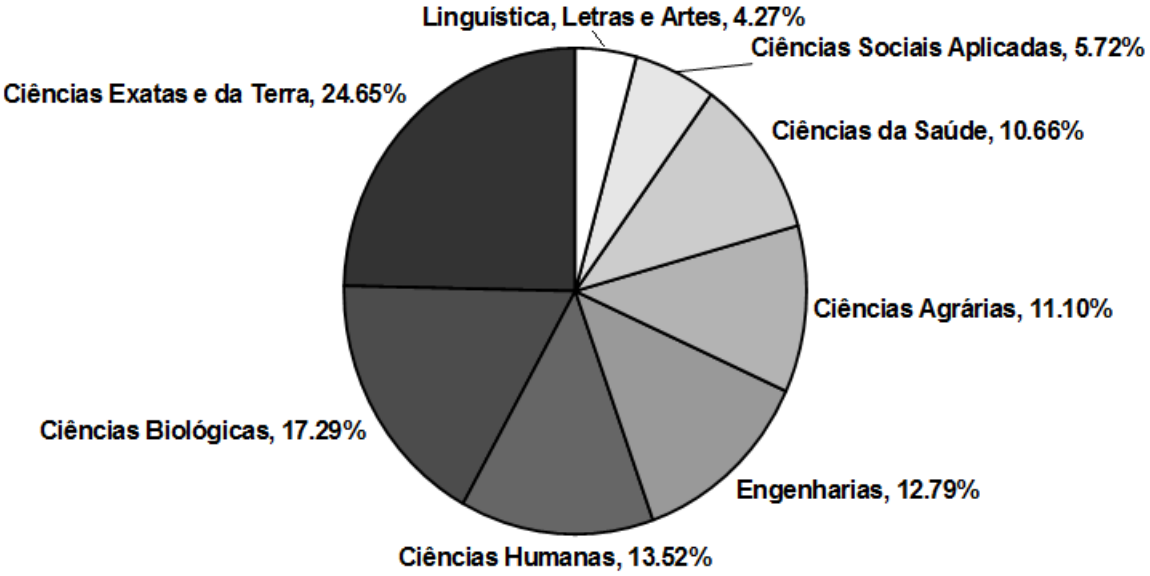
7.1.2 Apresentação e Análise dos Resultados

Nesta subseção são apresentados os resultados obtidos para a identificação de **Grandes Áreas**, **Áreas** e **Subáreas**.

Ao todo, 9.748 pesquisadores possuidores de bolsa produtividade haviam cadastrado apenas uma **Grande Área** de atuação. Seus currículos possuem, de 2001 a 2010, o registro de 300.756 artigos em periódicos, cujos títulos contêm 2.660.772 palavras (após a remoção das *stop-words*) e cada "palavra" diferente (após a execução do *stemmer*) apareceu, em média, 26,3 vezes nos registros das publicações (MIYATA; KANO; DIGIAMPIETRI, 2013a).

A figura 27 contém a distribuição destes 9.748 pesquisadores de acordo com sua grande-área de atuação. Observa-se que a grande-área mais frequente é "Ciências Exatas e da Terra", com 24,65% dos pesquisadores.

Figura 27 – Distribuição das grandes áreas na amostra



Fonte: adaptado de Miyata, Kano e Digiampietri (2013a)

A tabela 28 apresenta o resultado da identificação das grandes áreas utilizando-se apenas mineração de textos. Além de analisar a capacidade da característica utilizada na identificação, estes dados também pretendem verificar a influência temporal destes resultados. Cada linha da tabela indica a quantidade de anos utilizada nos dados de

treinamento (de um ano, dados apenas de 2010, a dez anos, dados de 2001 a 2010). As colunas indicam o período utilizado nos testes. Observa-se uma melhoria nos resultados à medida que dados de mais anos são utilizados, especialmente em relação ao período dos dados de teste. O melhor resultado, destacado na tabela, foi de 86,67% de acerto utilizando-se as publicações de nove anos (de 2002 a 2010) para os conjuntos tanto de treinamento quanto de testes.

Tabela 28 – Taxas de acerto utilizando mineração de textos - Grandes Áreas

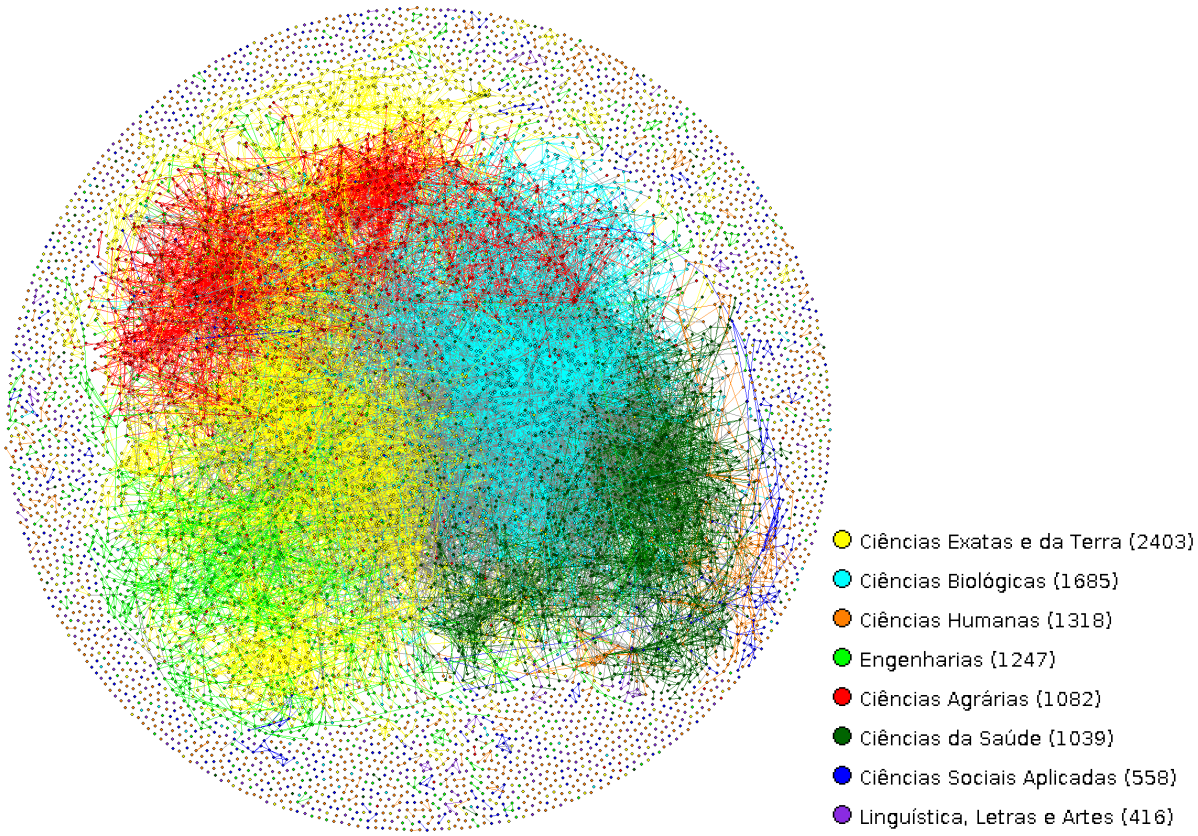
	1 ano	2 anos	3 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos	10 anos
1 ano	66,15%	75,79%	77,95%	80,41%	81,13%	82,36%	82,67%	82,77%	82,77%	83,08%
2 anos	66,87%	76,92%	79,59%	81,74%	81,85%	82,56%	82,97%	83,79%	84,21%	84,72%
3 anos	66,87%	76,92%	80,51%	81,54%	82,15%	82,46%	82,77%	82,87%	83,69%	83,49%
4 anos	66,56%	76,31%	80,41%	81,23%	82,87%	82,36%	82,67%	83,69%	84,31%	84,41%
5 anos	65,95%	76,72%	80,92%	81,64%	83,49%	83,18%	83,59%	84,72%	85,23%	85,03%
6 anos	66,26%	76,92%	80,72%	81,74%	83,59%	84,00%	83,49%	84,10%	84,41%	84,82%
7 anos	66,36%	77,03%	80,21%	81,03%	82,87%	84,10%	84,82%	85,23%	85,54%	85,33%
8 anos	66,77%	77,85%	80,82%	81,44%	83,38%	84,82%	85,13%	85,33%	86,15%	85,85%
9 anos	66,97%	77,85%	80,31%	81,23%	83,28%	85,03%	85,64%	85,95%	**86,67%**	86,46%
10 anos	67,08%	77,23%	80,62%	81,44%	83,18%	84,51%	84,92%	85,03%	86,36%	86,05%

Fonte: Miyata, Kano e Digiampietri (2013a)

As duas características oriundas da análise de redes sociais, isto é, vizinhança nível um (V1) e vizinhança nível dois (V2), também foram analisadas considerando-se dados de diferentes períodos de tempo. Para isto, dez grafos de coautoria foram criados (utilizando-se dados de 1 a 10 anos). A figura 28 contém a rede de coautorias considerando-se as publicações de 2001 até 2010. Nesta rede, todos os nós estão coloridos conforme sua área de atuação, inclusive os nós do conjunto de testes, pois o objetivo da figura é ilustrar como nós de uma mesma área, de um modo geral, estão agrupados. Um fenômeno parecido já havia sido destacado nas redes sociais apresentadas em outros capítulos.

Outro destaque da figura 28 é que há diversos nós isolados (sem ligação com nenhum outro nó). Estes nós não podem ter suas áreas de atuação identificadas utilizando-se as características baseadas na análise de redes sociais. Um problema semelhante ocorreu com o uso da característica baseada em mineração de textos, mas, neste último caso, apenas para os pesquisadores que não haviam publicado nenhum artigo no período considerado. A tabela 29 apresenta a porcentagem de pesquisadores cuja identificação não foi possível devido a estes fatores, observa-se que a técnica baseada em mineração de textos é mais inclusiva. Em todos os dados apresentados nesta seção, os pesquisadores que não puderam ter suas áreas identificadas foram contabilizados como erros de identificação (já que a

Figura 28 – Rede de coautorias - publicações de 2001 a 2010 - Grandes Áreas



Fonte: Miyata, Kano e Digiampietri (2013a)

identificação correta não foi possível). Observa-se que, utilizando dados de apenas um ano, isto é, de 2010, quase metade dos pesquisadores não poderiam ter sua grande-área de atuação identificada utilizando-se as características de ARS.

Tabela 29 – Pesquisadores que não puderam ser classificados - Grandes Áreas

	1 ano	2 anos	3 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos	10 anos
MT	13,03%	3,79%	1,44%	0,51%	0,41%	0,21%	0,21%	0,21%	0,21%	0,10%
V1 e V2	48,10%	34,46%	29,33%	26,05%	24,10%	22,77%	21,03%	20,10%	19,49%	18,77%

Fonte: Miyata, Kano e Digiampietri (2013a)

As taxas de acerto do uso de ARS e MT e de suas combinações são apresentadas na tabela 30. Para as combinações, utilizou-se a soma simples dos valores dos vetores de características, e cada pesquisador foi classificado de acordo com a grande área cujo valor apresentava a maior soma. As colunas contêm os períodos utilizados para treinamento e teste. As linhas apresentam as características e suas combinações (MIYATA; KANO; DIGIAMPIETRI, 2013a).

Tabela 30 – Resultados da combinação das técnicas para Grandes Áreas

	1 ano	2 anos	3 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos	10 anos
MT	66,15%	76,92%	80,51%	81,23%	83,49%	84,00%	84,82%	85,33%	86,67%	86,05%
V1	45,44%	57,44%	62,46%	66,77%	68,31%	69,23%	70,67%	71,69%	72,31%	73,03%
V2	48,10%	59,49%	64,21%	67,59%	69,23%	69,64%	70,67%	71,18%	71,59%	71,90%
MT+V1+V2	70,67%	80,41%	85,44%	85,85%	87,69%	88,41%	88,82%	89,03%	89,85%	89,54%
MT+V1	70,26%	80,21%	85,85%	86,05%	88,00%	88,92%	88,72%	89,23%	89,85%	89,44%
MT+V2	72,00%	81,85%	86,36%	87,08%	88,31%	89,03%	89,33%	89,95%	**90,56%**	90,26%
V1+V2	46,77%	58,56%	63,59%	67,49%	68,92%	69,44%	70,97%	71,79%	72,41%	72,92%

Fonte: Miyata, Kano e Digiampietri (2013a)

Individualmente, a característica baseada em MT obteve melhores resultados para todos os períodos analisados. Já a combinação com melhores resultados foi a de MT com V2, atingindo uma taxa de acerto de 90,56% para o período de treinamento e testes de 9 anos (destacado na tabela).

A tabela 31 contém a matriz de confusão da combinação de MT com V2 para o período de 10 anos (cuja taxa de acerto foi de 90,26%). Observa-se para a maioria das linhas que os erros de identificação mais frequentes ocorrem entre áreas consideradas próximas (por exemplo, na primeira linha "Ciência Agrárias" e "Ciências Biológicas", na segunda, "Ciências Biológicas" e "Ciências da Saúde", e na terceira "Ciências Exatas e da Terra" e "Engenharias").

Tabela 31 – Matriz de confusão - resultados utilizando MT combinada com V2

	Ciências Agrárias	Ciências Biológicas	Ciências Exatas e da Terra	Ciências Humanas	Ciências Sociais Aplicadas	Ciências da Saúde	Engenharias	Linguística, Letras e Artes	Não Classificado	Total
Ciências Agrárias	96	6	0	0	0	0	0	0	0	**102**
Ciências Biológicas	3	172	1	0	0	4	1	0	0	**181**
Ciências Exatas e da Terra	1	2	216	0	0	2	6	0	0	**227**
Ciências Humanas	0	2	1	110	9	4	0	10	0	**136**
Ciências Sociais Aplicadas	0	0	0	5	41	0	1	2	0	**49**
Ciências da Saúde	0	10	0	5	0	88	0	0	0	**103**
Engenharias	3	1	7	1	6	0	115	0	0	**133**
Linguística, Letras e Artes	0	0	0	0	1	0	0	42	1	**44**
Total	**103**	**193**	**225**	**121**	**57**	**98**	**123**	**54**	**1**	**975**

Fonte: Miyata, Kano e Digiampietri (2013a)

Abordagens mais complexas e/ou robustas podem ser utilizadas para a combinação das técnicas. O uso destas abordagens fugiu do escopo deste trabalho, mas, apenas para ilustrar, o algoritmo de inteligência artificial *Rotation Forest* obteve resultados levemente

melhores do que a combinação simples, sendo que o melhor resultado ocorreu para o período de 10 anos e a taxa de acerto foi de 90,87%.

Da amostra utilizada, 7.297 pesquisadores declaram atuar em apenas uma **área**, dentre as 76 encontradas na amostra. A área mais frequente foi "Física" contendo 8,74% dos pesquisadores. A tabela 32 apresenta a porcentagem de pesquisadores que não puderam ser classificados. A tabela 33 contém as taxas de acerto utilizando-se cada uma das características e suas combinações para os diferentes períodos de treinamento e teste.

Tabela 32 – Pesquisadores que não puderam ser classificados - Áreas

	1 ano	2 anos	3 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos	10 anos
MT	13,42%	3,42%	1,37%	0,55%	0,41%	0,27%	0,27%	0,27%	0,27%	0,14%
V1 e V2	54,79%	41,51%	35,75%	32,05%	29,73%	28,49%	26,03%	24,93%	24,38%	23,84%

Fonte: Miyata, Kano e Digiampietri (2013a)

Para a identificação de áreas, a característica baseada em MT só obteve os melhores resultados individuais para períodos envolvendo quatro ou mais anos. Para períodos menores, a característica V2 obteve os melhores resultados. Novamente, a combinação de MT e V2 foi a combinação que obteve os melhores resultados (84,11% para o período de 10 anos de treinamento e testes).

Tabela 33 – Resultados da combinação das técnicas para Áreas

	1 ano	2 anos	3 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos	10 anos
MT	36,58%	48,49%	53,56%	59,45%	63,15%	64,11%	65,62%	66,99%	67,26%	68,22%
V1	37,12%	46,30%	51,78%	55,34%	57,67%	58,77%	61,10%	61,51%	61,51%	61,64%
V2	40,00%	49,73%	55,34%	58,49%	60,55%	61,10%	62,88%	63,42%	63,84%	63,70%
MT+V1+V2	51,64%	62,05%	66,85%	73,70%	76,85%	78,22%	79,45%	79,18%	80,14%	80,96%
MT+V1	50,00%	61,23%	66,58%	72,47%	76,03%	76,58%	78,49%	77,81%	78,49%	79,18%
MT+V2	52,33%	64,38%	68,63%	75,07%	78,77%	80,27%	81,92%	81,78%	83,29%	**84,11%**
V1+V2	37,95%	46,71%	52,60%	56,99%	60,00%	60,68%	62,05%	62,88%	63,01%	63,01%

Fonte: Miyata, Kano e Digiampietri (2013a)

Da amostra, apenas 3.427 pesquisadores declaram atuar em apenas uma **subárea** (dentre as 443 subáreas declaradas). A subárea mais frequente foi "Física da Matéria Condensada", com 6,92% dos pesquisadores (MIYATA; KANO; DIGIAMPIETRI, 2013a).

A tabela 34 apresenta a porcentagem de pesquisadores que não puderam ter suas subáreas identificadas. Observa-se que, utilizando-se V1 ou V2, não seria possível essa identificação para quase dois terços destes pesquisadores, considerando os dados de apenas um ano, e de mais de 30% para os dados de 10 anos.

Tabela 34 – Pesquisadores que não puderam ser classificados - Subáreas

	1 ano	2 anos	3 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos	10 anos
MT	11,66%	3,21%	1,17%	0,87%	0,58%	0,58%	5,25%	0,58%	0,58%	0,29%
V1 e V2	64,14%	53,35%	45,77%	39,94%	38,48%	35,86%	33,24%	31,49%	30,61%	30,03%

Fonte: Miyata, Kano e Digiampietri (2013a)

A tabela 35 contém as taxas de acerto utilizando-se as diferentes características e suas combinações. A característica baseada em MT teve os piores resultados em todos os períodos. Pode-se concluir que esta técnica é mais sensível ao volume de dados disponível para treinamento, assim, conjuntos de dados mais esparsos (entre as centenas de subáreas) influenciaram os resultados. A técnica V2 foi novamente melhor do que a V1 e, novamente, a combinação de MT e V2 apresentou os melhores resultados. Destaca-se o acerto de 59,77% ocorrido para os períodos de 9 e 10 anos.

Tabela 35 – Resultados da combinação das técnicas para Subáreas

	1 ano	2 anos	3 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos	10 anos
MT	10,79%	14,87%	17,20%	20,41%	20,70%	23,03%	23,62%	26,53%	26,53%	26,24%
V1	26,53%	32,36%	38,78%	43,44%	44,31%	46,94%	48,69%	48,10%	48,10%	47,81%
V2	30,32%	37,61%	45,48%	48,98%	50,44%	52,19%	53,64%	54,81%	55,98%	56,27%
MT+V1+V2	31,20%	38,78%	43,44%	48,69%	50,44%	53,35%	55,39%	57,14%	55,98%	55,69%
MT+V1	29,15%	37,32%	41,98%	46,06%	47,52%	50,15%	52,48%	53,64%	52,19%	51,90%
MT+V2	32,07%	41,11%	46,65%	51,31%	53,94%	55,39%	58,02%	59,48%	**59,77%**	**59,77%**
V1+V2	27,11%	33,82%	39,36%	44,61%	46,06%	48,10%	49,85%	50,44%	50,15%	50,44%

Fonte: Miyata, Kano e Digiampietri (2013a)

Devido à grande quantidade de subáreas, o acerto de identificação da subárea de quase 60% dos pesquisadores foi considerado satisfatório.

7.1.3 Conclusões - identificação de áreas

Nesta seção foram apresentados resultados utilizando características simples oriundas da mineração de textos e análise de redes sociais para a identificação da área de atuação de pesquisadores.

Os resultados obtidos foram considerados satisfatórios, atingindo taxas de acerto superiores a 90% para a identificação das grandes áreas; superiores a 84% para a identificação de áreas; e de 59,77% para a identificação de subáreas (MIYATA; KANO; DIGIAMPIETRI, 2013a). Observou-se que a característica baseada em mineração de textos é mais sensível a quantidade de dados utilizados no treinamento, especialmente quando o número de áreas

(ou subáreas) é grande. Porém, é a característica mais inclusiva por não necessitar da existência de colaborações de coautoria entre os pesquisadores da amostra.

Apesar da observação da influência dos períodos de tempo utilizados no treinamento e teste não foi possível concluir qual o período ideal para este tipo de análise. Para isto, provavelmente seria necessário analisar janelas de tempo mais longas.

7.2 Análise de tendências

Neste trabalho a análise de tendências está sendo utilizada tanto para auxiliar na caracterização de grupos de pesquisa quanto para tentar predizer o comportamento futuro de alguns temas de pesquisa.

A premissa do trabalho é que combinar a análise de tendências baseada apenas em séries temporais com informações sobre os elementos geradores da informação (utilizando informações da análise de redes sociais) permite a identificação de tendências mais precisa possibilitando uma previsão mais acurada (TRUCOLO, 2015).

Para isto, optou-se por utilizar TF-IDF, mesma métrica utilizada na mineração de textos da seção anterior, como medida a ser analisada e predita.

7.2.1 Metodologia

O trabalho desenvolvido neste assunto foi iniciado por uma revisão da literatura sobre o tema, sumarizada no capítulo 2, seguida por três atividades: obtenção dos dados; extração automática de termos; e análise de tendências.

Obtenção dos dados. Assim como apresentado em outras seções, foram utilizados dados dos professores permanentes dos 45 programas de pós-graduação em Ciência da Computação avaliados no triênio 2007-2009 e que possuíam doutorado ou mestrado acadêmico. Foram identificados 57.501 títulos de publicações dos 889 professores no período de 1991 e 2011. Os dados de 1991 a 2010 (20 anos) foram utilizados para treinamento e o objetivo foi tentar predizer o comportamento de alguns termos em 2011.

Extração automática de termos. Todos os títulos das publicações foram pré-processadas com a remoção de *stop-words*. A fim de se identificar os termos mais "importantes" de um conjunto de documentos, foi utilizada a frequência adjacente das palavras que compõem esses termos, conforme apresentado por Nakagawa e Mori (2002).

Análise de tendências. A medida TF-IDF foi calculada para os termos extraídos. Esta medida foi utilizada tanto para distinguir/caracterizar a produção dos diferentes programas de pós-graduação como também tentou-se prever o valor desta medida com base no valor da medida nos anos anteriores, combinada com informações oriundas da análise da rede social acadêmica. Para a predição de valores foram verificadas diferentes funções para identificar aquela que mais se aproximava às medidas TF-IDF relacionadas a cada termo. Foram testadas a regressão linear, exponencial, logarítmica, do tipo *power law* e polinomial de grau 2 a 5. A regressão que foi considerada a que melhor representa um termo foi a que obteve o menor erro quadrático médio da função em relação aos valores medidos. Um termo foi considerado uma tendência se possuísse uma alta previsão da medida TD-IDF em relação aos demais termos.

7.2.2 Resultados

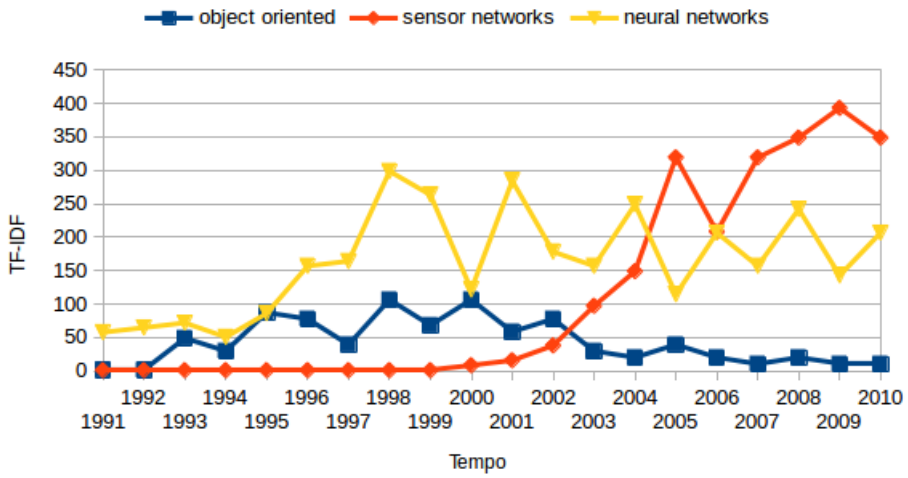
Esta seção é iniciada com resultados já publicados (TRUCOLO; DIGIAMPIETRI, 2014a) considerando apenas o uso da análise de séries temporais de documentos (no caso, títulos de publicações). Por fim, alguns resultados ainda não publicados são apresentados, utilizando também métricas oriundas da análise de redes sociais.

A figura 29 ilustra o comportamento temporal da medida TF-IDF de três termos extraídos. Estes termos foram escolhidos por apresentarem comportamentos bastante distintos. O termo *neural networks* possui esta medida relativamente alta ao longo do período com alguns altos e baixos; *sensor networks* vem se destacando recentemente; e, por fim, *object oriented* teve uma diminuição da medida TF-IDF nos últimos anos.

Também de maneira ilustrativa, as figuras 30 e 31 mostram a curva considerada mais adequada para os valores TF-IDF de dois termos. Na figura 30 observa-se a regressão polinomial de grau 3 e os valores TF-IDF do termo *sensor networks*. Já a figura 31 contém uma curva do tipo *power law* e o termo *object oriented*.

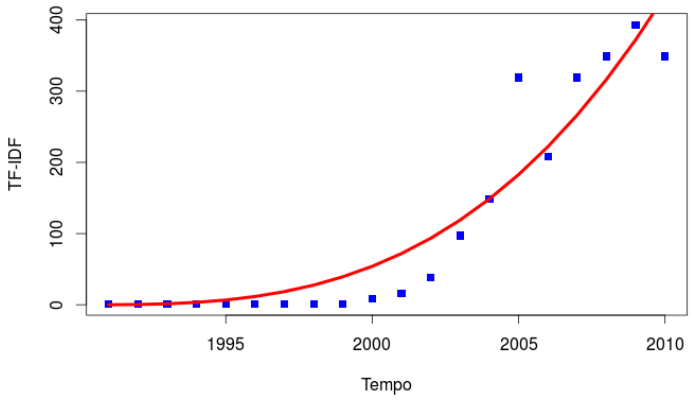
A tabela 36 contém os valores TF-IDF das que foram consideradas as 20 maiores tendências para o ano de 2011 (neste caso, os 20 maiores aumentos preditos para os valores de TF-IDF dos termos extraídos). O erro absoluto médio entre o valor predito e o valor real foi de aproximadamente 26,5% e a correlação entre estes valores foi de 0,68.

Figura 29 – Comportamento temporal de três termos



Fonte: Trucolo e Digiampietri (2014a)

Figura 30 – Curva de tendência gerada pela regressão não linear *power law* para o termo *sensor networks*



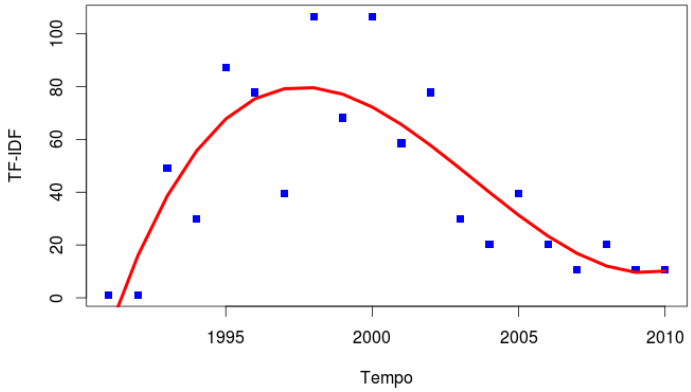
Fonte: Trucolo e Digiampietri (2014a)

Observa-se na tabela 36 a presença de um único termo em português *realidade aumentada*, lembrando-se que para este estudo não foi realizada a tradução de termos ou o filtro de títulos de apenas um idioma.

A análise de tendências utilizando-se a mesma medida foi aplicada também a cada um dos programas de pós-graduação. A tabela 37 contém as duas principais tendências de cada programa analisado. Observa-se tendências bastante variadas entre os programas.

Todos os resultados apresentados até o momento nesta seção consideraram apenas a série temporal dos títulos dos artigos. Em uma pesquisa ainda em desenvolvimento, o

Figura 31 – Curva de tendência gerada pela regressão não linear polinomial de grau 3 para o termo *object oriented*



Fonte: Trucolo e Digiampietri (2014a)

aluno de mestrado Caio Cesar Trucolo, sob a orientação do autor do presente documento, está avaliando o impacto do uso de métricas da análise de redes sociais na previsão de tendências. A premissa desse estudo é que a importância (ou influência) de quem gera a informação (no caso, produz um artigo científico) irá influenciar no fato de o assunto (ou dos termos) relacionado ao artigo se tornar ou não uma tendência.

Para isto, foi selecionada a mesma amostra utilizada nos resultados apresentados nesta seção (dados dos docentes permanentes dos 45 programas), porém uma versão mais atual dos currículos foi copiada e considerou-se adequada a utilização das publicações até 2012 (treinamento até 2011 e testes usando os dados de 2012).

A figura 32 apresenta a rede de coautorias em artigos completos publicados entre os docentes dos 45 programas analisados, a numeração apresentada na legenda corresponde à mesma numeração utilizada nas tabela 37. Nesta figura, cada nó corresponde a um docente e cada aresta indica a presença de uma ou mais colaborações em publicações de artigos completos no período analisado. Os nós estão coloridos de acordo com o programa de pós-graduação. As arestas entre docentes de um mesmo programa possuem a mesma cor dos nós, já as arestas entre docentes de diferentes programas são da cor cinza.

Neste estudo (TRUCOLO, 2015), observou-se que a inclusão de medidas oriundas da análise de redes sociais permitiu uma predição com um erro, em média, ao menos 30% menor do que o erro obtido utilizando-se apenas a análise temporal dos documentos. A tabela 38 apresenta os valores reais de TF-IDF para o ano de 2012, o valor predito utilizando-se apenas a série histórica de valores TF-IDF e o valor predito utilizando-se

Tabela 36 – Principais tendências em relação aos termos extraídos

Termo	Previsão TF-IDF para 2011
product line	413,57
wireless sensor	402,99
sensor networks	321,47
wireless sensor networks	320,69
neural networks	277,29
software product	255,62
product lines	243,05
software development	238,76
particle swarm optimization	227,22
swarm optimization	227,22
particle swarm	224,63
optimum-path forest	219,73
realidade aumentada	209,51
time series	208,92
genetic algorithm	207,47
case study	204,01
scheduling problem	181,71
social networks	181,41
infocomp ufla	176,29
genetic programming	173,48

Fonte: Trucolo e Digiampietri (2014a)

também análise de redes sociais. Enquanto a combinação das diferentes métricas oriundas da análise de redes sociais ainda está em teste, os resultados apresentados nesta tabela foram obtidos pela seguinte fórmula aplicada a cada termo extraído:

$$TFIDF_{predito} = 0,42 + 0,48 * NosComp + 0,52 * AutoCent + 0,31 * RST$$

sendo, $TFIDF_{predito}$ o valor a ser predito da medida TF-IDF; $NosComp$ corresponde ao número de nós do componente no qual o artigo foi publicado; $AutoCent$ é a centralidade de autovalor média dos nós do componente; RST corresponde à regressão utilizando apenas série temporal (conforme apresentado anteriormente)

Pela tabela 38, fica evidente o ganho de precisão ao se comparar os dois modelos. O erro gerado pelo modelo com fator de redes sociais corresponde a apenas 17% do erro total gerado pelo modelo de regressão simples.

Tabela 37 – Principais tendências de termos em cada programa

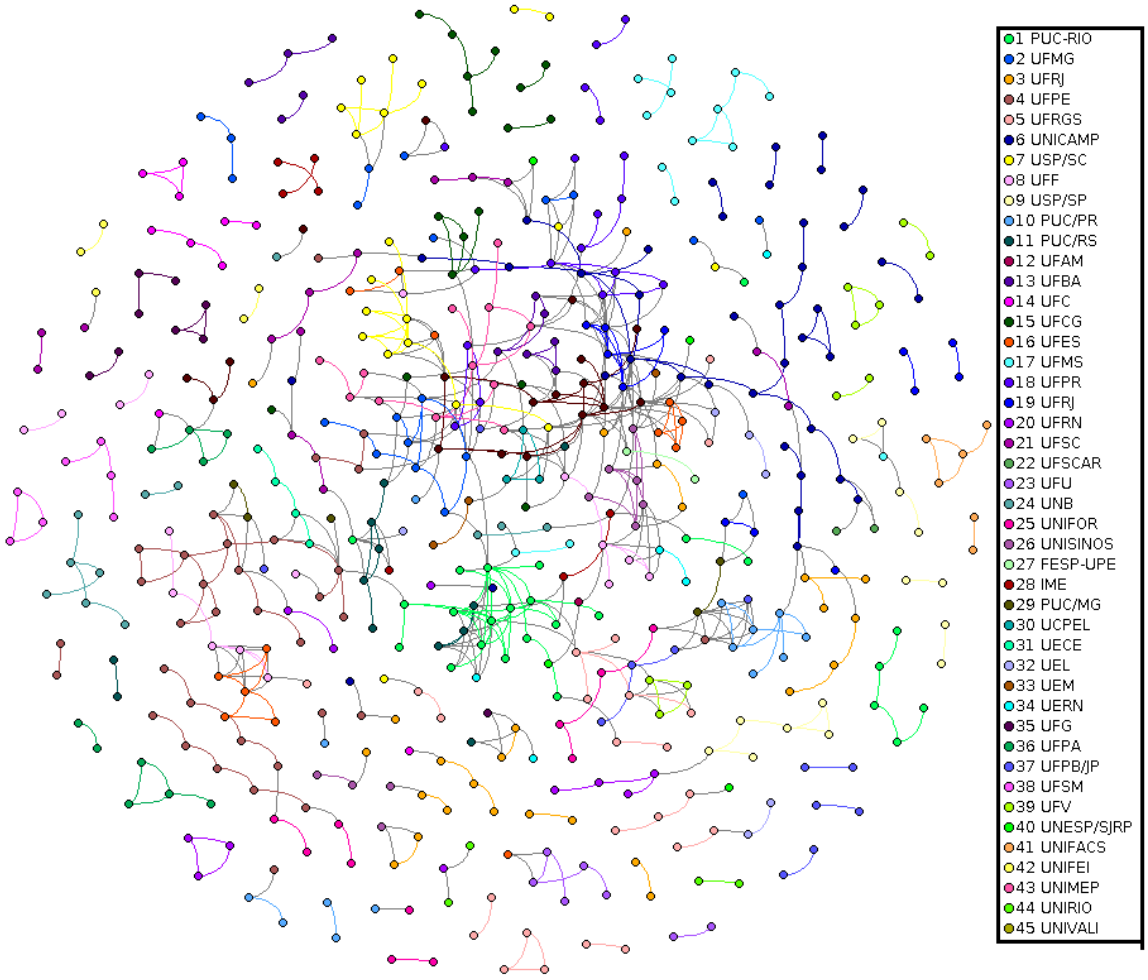
	Programa	Primeira tendência	Segunda tendência
1	PUC-RIO - Informática	product line	microscopy images
2	UFMG - Ciências da Computação	genetic programming	name disambiguation
3	UFRJ - Engenharia de Sistemas e Computação	hyperbolic smoothing	clustering method
4	UFPE - Ciências da Computação	software development	time series
5	UFRGS - Computação	sensor networks	eye fundus images
6	UNICAMP - Ciência da Computação	optimum-path forest	foresting transform
7	USP / SC - Ciências da Computação e Matemática Computacional	neural networks	time series
8	UFF - Computação	wave propagation	cellular automata
9	USP - Ciências da Computação	oriented relational	field-research oriented relational database
10	PUC / PR - Informática	arq scheme	music genre
11	PUC / RS - Ciência da Computação	promising protein receptor snapshots	infocomp (ufla
12	UFAM - Informática	grounded theory	medical devices
13	UFBA - Ciência da Computação - UFBA - UNIFACS	real-time systems	failure detectors
14	UFC - Ciências da Computação	software product	data integration
15	UFCG - Ciência da Computação	power management	application user interfaces
16	UFES - Informática	fault diagnosis	matrices reordering algorithms
17	UFMS - Ciência da Computação	shuffling experiments	e-sapi bovis
18	UFPR - Informática	swarm optimization	particle swarm optimization
19	UFRJ - Informática	três empresários protagonistas	deficientes visuais
20	UFRN - Sistemas e Computação	product line	wireless sensor
21	UFSC - Ciências da Computação	image segmentation	crônicas não-transmissíveis
22	UFSCAR - Ciência da Computação	production scheduling	engenharia elétrica
23	UFU - Ciência da Computação	trace alignment	diferenciais parciais
24	UNB - Informática	particle swarm optimization	particle swarm
25	UNIFOR - Informática Aplicada	wireless sensor	decision analysis
26	UNISINOS - Computação Aplicada	multilevel approach	composite structure
27	FESP / UPE - Engenharia da Computação	time series	particle swarm optimization
28	IME - Sistemas e Computação	upper bound	web services
29	PUC / MG - Informática	sensor networks	graph matching
30	UCPEL - Informática	simulação quântica	architecture using
31	UECE - Ciência da Computação	test case	release planning
32	UEL - Ciência da Computação	spectral analysis	detection using
33	UEM - Ciência da Computação	users personomy	personomy using
34	UERN - Ciência da Computação - UERN - UFERSA	sensor networks	transcoded videos
35	UFG - Ciência da Computação	capacitated arc	capacitated arc routing
36	UFPA - Ciência da Computação	iso/iec 12207	neurais artificiais
37	UFPB / J.P. - Informática	naive bayes	training assessment
38	UFSM - Informática	estilos cognitivos	hipermídia adaptativa
39	UFV - Ciência da Computação	scheduling problem	sequence dependent setup
40	UNESP / SJRP - Ciência da Computação	optimum-path forest	contours initialized
41	UNIFACS - Sistemas e Computação	developers context-specific preferred representational	preferred representational
42	UNIFEI - Ciência e Tecnologia da Computação	rough sets	self-organizing map model
43	UNIMEP - Ciência da Computação	realidade aumentada	augmented reality
44	UNIRIO - Informática	business models	case study
45	UNIVALI - Computação	process capability models	neurais artificiais

Fonte: Trucolo e Digiampietri (2014a)

7.2.3 Considerações finais

Este trabalho apresentou informações gerais sobre as tendências da produção científica brasileira da área de ciência da computação utilizando uma forma automática de extração de termos e expressões. Foi realizada uma análise de tendências geral e análises

Figura 32 – Redes de coautoria dos programas analisados



Fonte: Trucolo e Digiampietri (2014a)

individuais para cada programa. Dessa forma, foi possível identificar quais assuntos estão em alta e quais estão em baixa. Além disso, foi realizada uma análise da rede de coautorias dos docentes-orientadores nos programas de pós-graduação da área de ciência da computação na qual foi observado que não há forte correlação entre as tendências encontradas nos programas e as coautorias entre programas.

Os resultados deste trabalho consistem em um passo inicial considerando-se todo o potencial da análise de tendência da produção científica nacional.

Adicionalmente, pretende-se agrupar os docentes e os termos de acordo com as subáreas da Ciência da Computação a fim de se identificar a dinâmica e as tendências nas publicações nestas subáreas.

Tabela 38 – Resultados preliminares da previsão da medida TD-IDF para o ano de 2012

Termo	**Real**	**Série Temporal**	**Erro**	**Proposta**	**Erro**
service discovery	135,17	441,52	306,35	123,39	11,77
information systems	147,32	334,29	186,97	148,37	1,05
supply chain	174,31	298,37	124,06	143,96	30,35
web services	225,28	297,74	72,46	201,05	24,23
product line	174,99	291,57	116,57	154,73	20,26
motion estimation	107,78	274,36	166,58	99,00	8,78
social network	249,05	269,42	20,38	198,94	50,11
business process	131,75	240,09	108,34	119,61	12,14
time series	150,79	217,76	66,97	147,03	3,76
neural network	213,36	178,86	34,51	198,85	14,51

Fonte: Adaptado de Trucolo (2015)

7.3 Relação orientador-orientado

Nas redes sociais acadêmicas, duas das relações mais estudadas são a de colaboração na publicação de artigos e a relação de orientação. Participar do desenvolvimento de um projeto, da escrita de um artigo, de sua submissão e eventual publicação são atividades muito importantes na vida acadêmica. Por isto, muitos orientadores têm incentivado que seus alunos, mesmo os de iniciação científica ou trabalho de conclusão de curso, participem de todas estas atividades.

Por outro lado, o orientado, além de aprender, também pode (ou deve) desempenhar um papel ativo no desenvolvimento de seu projeto, auxiliando muitas vezes em diferentes pesquisas de seu orientador.

Nesta seção são apresentados alguns resultados sobre uma análise realizada sobre a participação dos orientados na produção científica de seus orientadores dentro da área de Ciência da Computação.

7.3.1 Metodologia

O trabalho desenvolvido foi composto por três atividades: seleção da amostra; identificação das informações relevantes; e análise dos dados.

Seleção da amostra. A mostra selecionada para este trabalho foi baseada nos 889 professores permanentes dos programas de pós-graduação em Ciência da Computação da CAPES que foram avaliados no triênio 2007-2009. Estes professores serão as pessoas

denominadas de “orientadores”, além deles, fazem parte da amostra todos os seus “orientados” cujos currículos puderam ser encontrados, bem como todas as demais pessoas cujo currículo esteja explicitamente referenciado dentro dos currículos dos 889 professores (*links explícitos*). Estas pessoas serão denominadas “vizinhos”.

Identificação das informações relevantes. Para este estudo, as informações utilizadas foram: todos currículos relacionados a cada orientador; identificação dos orientandos; identificação das produções bibliográfica dos orientadores; identificação das coautorias; e identificação dos primeiros autores de cada produção bibliográfica.

Para a **identificação dos currículos relacionados a cada pesquisador** foram utilizadas as relações explícitas de cada currículo. Isto é, os *links* HTML existentes nos currículos dos orientadores e que podem se referenciar, por exemplo, a currículos de orientados ou de coautores. Esta atividade identificou 12.843 novos currículos.

Para a **identificação dos currículos dos orientados**, alguns dos registros das orientações já possuíam um *link* para o currículo do orientado. Para os demais, foram comparados o nome completo do possuidor de cada currículo da área de Ciência da Computação com o nome presente na lista de orientações. Desta forma, 6.265 currículos de orientados foram identificados.

Ao todo, 16.403 currículos foram analisados neste estudo (correspondendo aos currículos dos orientadores, orientados e currículos relacionados [*vizinhos*]). Um banco de dados relacional (ver capítulo 3) foi criado utilizando-se os dados destes currículos e as relações de coautoria foram identificadas (ver seção 4.1).

A verificação da **presença de orientados na lista de autores** das produção bibliográficas dos orientadores foi realizada utilizando-se a primeira estratégia de resolução de nomes de autores apresentada na seção 4.2. Esta estratégia utiliza um nome completo como referência e o procura na lista de autores de um artigo (cujos nomes tipicamente estão abreviados). O nome de cada orientado (extraído da lista de orientações do currículo do orientador) foi utilizado como referência e era comparado com os nomes dos coautores de todas as publicações do orientador. Este processo também foi utilizado para verificar se o primeiro autor de cada artigo era ou não o orientador ou um de seus orientados. Para esta atividade utilizou-se apenas o currículo do orientador, pois além de possuir todas as informações necessárias, um cruzamento com os possíveis currículos dos orientados poderia inserir ruídos na análise, pois nem todos os orientados possuem currículo Lattes e, potencialmente, nem todos os currículos estarão devidamente atualizados.

Análise dos dados. Com base nos dados obtidos e organizados nas tarefas anteriores foram realizadas algumas análises confrontando e correlacionando dados referentes à produção bibliográfica do orientador com dados de suas orientações. Adicionalmente, uma rede social de coautorias composta pelos orientadores, orientados e *vizinhos* foi produzida e analisada.

7.3.2 Resultados

Cada um dos orientadores analisados orientou, na média, mais de 44 alunos, conforme pode ser observado na tabela 39.

Tabela 39 – Orientações por tipo

	Supervisão de Pós-Doutorado	Tese de Doutorado	Dissertação de Mestrado	Iniciação Científica	Trabalho de Conclusão de Curso	Orientação de Outra Natureza	Monografia de Conclusão de Curso de Aperfeiçoamento ou Especialização	Orientações Totais
Total	194	2459	13929	7808	10500	2273	2086	39249
Média	0,22	2,77	15,67	8,78	11,81	2,56	2,35	44,15
Mediana	0	1	12	5	7	0	0	36
Desvio Padrão	0,85	4,93	14,19	11,08	15,07	14,18	7,72	34,69

Fonte: Digiampietri, Mugnaini e Alves (2013)

Para o grupo de orientadores analisados, as três modalidades mais frequentes de orientação são dissertação de mestrado, trabalho de conclusão de curso e iniciação científica.

A tabela 40 contém a porcentagem das publicações dos orientadores que possuem ao menos um orientado na lista de autores. Do total, mais de metade das publicações têm a participação dos orientados.

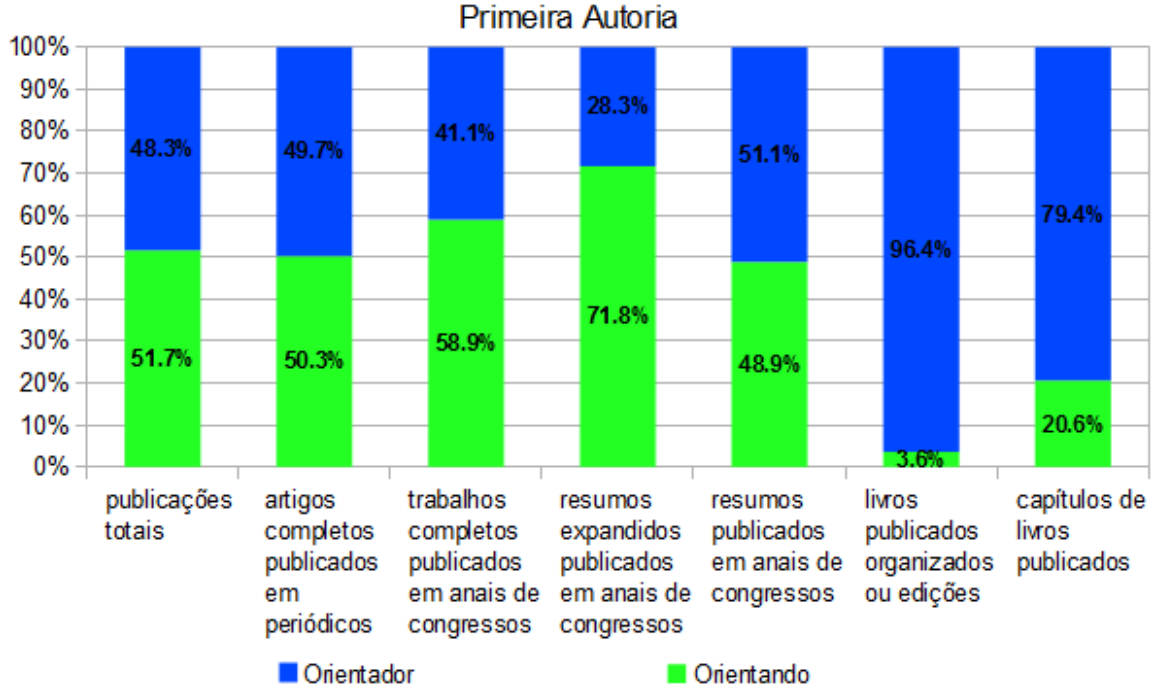
Ao se considerar apenas as publicações do orientador nas quais o primeiro autor é o próprio orientador ou um de seus orientados, observa-se que mais de metade possui um orientado como primeiro autor (figura 33). A maior porcentagem de primeira autoria de um dos orientados ocorre nos resumos expandidos (71,8%), no extremo oposto, o primeiro autor de livros costuma ser o orientador (96,4%).

Tabela 40 – Porcentagem de publicações com a participação de orientados

	artigos completos publicados em periódicos	trabalhos completos publicados em anais de congressos	resumos expandidos publicados em anais de congressos	resumos publicados em anais de congressos	livros publicados organizados ou edições	capítulos de livros publicados	total
publicações com orientandos envolvidos	47,50%	57,30%	61,60%	53,40%	6,20%	27,70%	51,60%

Fonte: Digiampietri, Mugnaini e Alves (2013)

Figura 33 – Distribuição dos primeiros autores entre orientadores e orientandos



Fonte: Digiampietri, Mugnaini e Alves (2013)

Neste trabalho também se computou a quantidade média de publicações de cada orientado em colaboração com seu orientador. É possível observar na tabela 41 que, na média, cada supervisão de pós-doutorado produz 1,64 artigos em periódicos e 3,7 trabalhos completos em anais.

A tabela 42 apresenta a porcentagem relativa dos dados da tabela 41 para cada tipo de orientação. Assim, a soma de cada linha desta tabela vale 100%. A partir destes dados é possível verificar qual o enfoque, em termos de tipo de publicação, tem sido dado para cada tipo de orientação. Por exemplo, observa-se que, proporcionalmente, as supervisões de pós-doutorado são as que possuem maior destaque na publicação de periódicos enquanto as dissertações de mestrado se destacam na produção de artigos completos em anais. Ao

Tabela 41 – Quantidade média de participações dos orientados nas publicações do orientador

	artigos completos publicados em periódicos	trabalhos completos publicados em anais de congressos	resumos expandidos publicados em anais de congressos	resumos publicados em anais de congressos	livros publicados organizados ou edições	capítulos de livros publicados	total de publicações com participação do orientando
supervisão de pós-doutorado	1,64	3,7	0,36	0,35	0,06	0,3	6,41
tese de doutorado	1,62	5,05	0,35	0,47	0,05	0,4	7,94
dissertação de mestrado	0,28	1,5	0,11	0,14	0,01	0,07	2,11
iniciação científica	0,08	0,47	0,08	0,24	0	0,02	0,88
trabalho de conclusão de curso	0,04	0,29	0,04	0,06	0	0,01	0,45
orientação de outra natureza	0,06	0,33	0,05	0,16	0	0,02	0,62
monografia de conclusão de curso de aperfeiçoamento ou especialização	0,03	0,1	0,01	0,02	0	0,01	0,17
média do orientador	**16,19**	**58,74**	**4,6**	**9,7**	**3,91**	**6,44**	**99,58**

Fonte: Digiampietri, Mugnaini e Alves (2013)

se analisar a coluna relativa a resumos expandidos, observa-se que são as orientações de iniciação científica que mais se destacam (possuem maior valor relativo nesta coluna).

Tabela 42 – Distribuição das participações dos orientados por tipo de produção

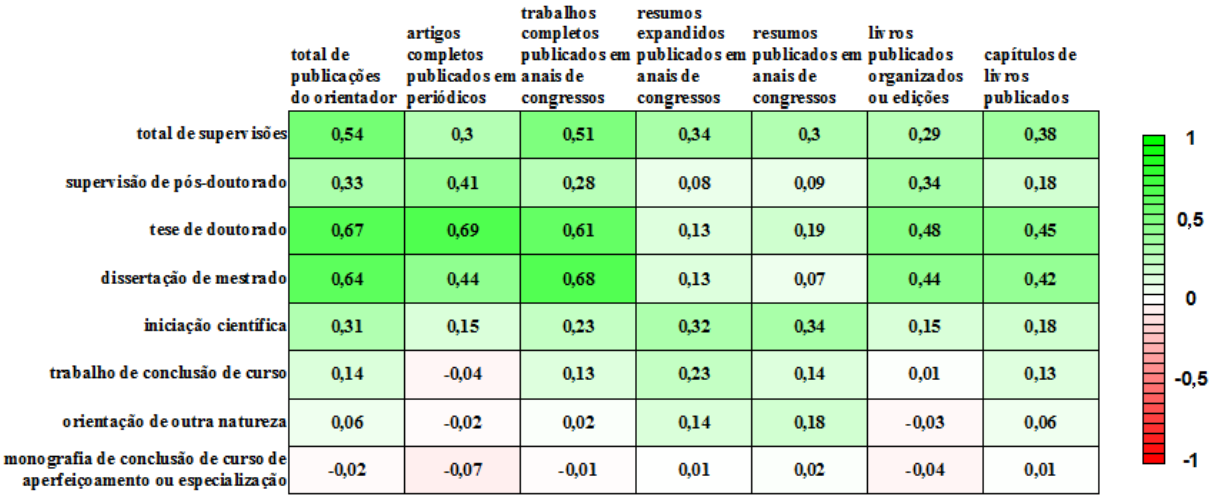
	artigos completos publicados em periódicos	trabalhos completos publicados em anais de congressos	resumos expandidos publicados em anais de congressos	resumos publicados em anais de congressos	livros publicados organizados ou edições	capítulos de livros publicados
supervisão de pós-doutorado	25,58%	57,76%	5,55%	5,47%	0,97%	4,67%
tese de doutorado	20,44%	63,63%	4,45%	5,86%	0,63%	4,99%
dissertação de mestrado	13,05%	71,29%	5,24%	6,80%	0,39%	3,22%
iniciação científica	8,71%	53,55%	8,80%	26,81%	0,12%	2,02%
trabalho de conclusão de curso	9,66%	64,37%	8,36%	14,09%	0,19%	3,33%
orientação de outra natureza	9,65%	53,42%	8,39%	25,09%	0,63%	2,82%
monografia de conclusão de curso de aperfeiçoamento ou especialização	18,38%	58,77%	7,80%	10,58%	0,28%	4,18%
média do orientador	**16,26%**	**58,99%**	**4,62%**	**9,74%**	**3,93%**	**6,47%**

Fonte: Digiampietri, Mugnaini e Alves (2013)

A figura 34 apresenta os valores de correlação de Pearson entre a quantidade de orientações e quantidade de publicações de cada autor. Há uma correlação superior a 0,5 entre o total de orientações e o total de publicações. Destacam-se os valores 0,69 entre o

número de orientações de doutorado e o total de artigos publicados em periódicos e 0,68 entre o número de orientações de mestrado e o total de artigos completos publicados em anais.

Figura 34 – Correlação entre a quantidade de supervisões e de produções



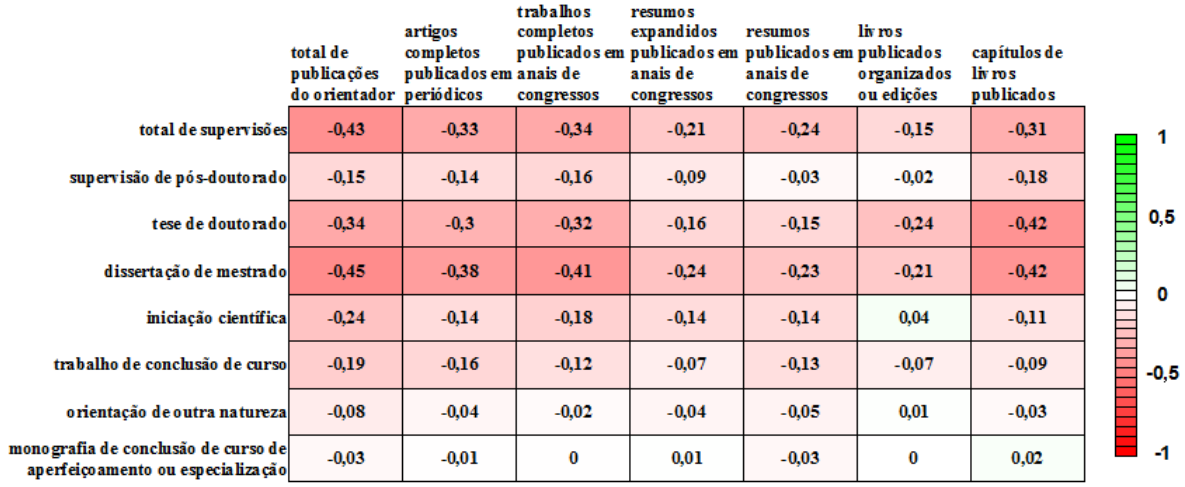
	total de publicações do orientador	artigos completos publicados em periódicos	trabalhos completos publicados em anais de congressos	resumos expandidos publicados em anais de congressos	resumos publicados em anais de congressos	livros publicados organizados ou edições	capítulos de livros publicados
total de supervisões	0,54	0,3	0,51	0,34	0,3	0,29	0,38
supervisão de pós-doutorado	0,33	0,41	0,28	0,08	0,09	0,34	0,18
tese de doutorado	0,67	0,69	0,61	0,13	0,19	0,48	0,45
dissertação de mestrado	0,64	0,44	0,68	0,13	0,07	0,44	0,42
iniciação científica	0,31	0,15	0,23	0,32	0,34	0,15	0,18
trabalho de conclusão de curso	0,14	-0,04	0,13	0,23	0,14	0,01	0,13
orientação de outra natureza	0,06	-0,02	0,02	0,14	0,18	-0,03	0,06
monografia de conclusão de curso de aperfeiçoamento ou especialização	-0,02	-0,07	-0,01	0,01	0,02	-0,04	0,01

Fonte: Digiampietri, Mugnaini e Alves (2013)

Também foram computadas as correlações entre o número de publicações tendo o orientador como primeiro autor e o número de orientações (figura 35). Uma correlação de -0,43 foi medida entre o número total de publicações em primeira autoria e o número total de orientações. Dentre as orientações, a que está mais correlacionada (negativamente) com o total de publicações em coautorias é a de dissertações de mestrado (-0,43). Uma possível justificativa para isto é o fato desse tipo de orientação ser bem mais frequente do que a de doutorado.

Catorze redes de coautorias foram produzidas e analisadas, compostas pelos orientadores, seus orientados e demais vizinhos: uma para cada ano de 2000 a 2012 e uma contendo as relações acumuladas do período. A figura 36 apresenta a rede que considerou as coautorias de todo o período. Cada nó representa uma pessoa, sendo que nós em azul representam orientadores, em verde representam orientados e em cinza as demais pessoas. As relações de coautoria são representadas por arestas. Arestas da cor vermelha indicam uma coautoria entre orientador e orientado, as demais arestas que envolvem orientador estão coloridas em azul e as demais arestas envolvendo orientados estão em verde. Por fim, arestas ligando duas pessoas que não sejam nem orientadores nem orientados estão representadas em cinza. O tamanho de cada nó da rede é proporcional ao valor da medida *Author Rank* (LIU et al., 2005) do respectivo nó.

Figura 35 – Correlação entre a quantidade de supervisões e a porcentagem de primeira autoria do orientador



Fonte: Digiampietri, Mugnaini e Alves (2013)

A figura 37 contém a variação dos valores médios de grau e *author rank* ao longo do tempo. Apesar das oscilações, observa-se que, ao longo do tempo, os orientados têm ganhado importância em relação aos demais colaboradores dos orientadores.

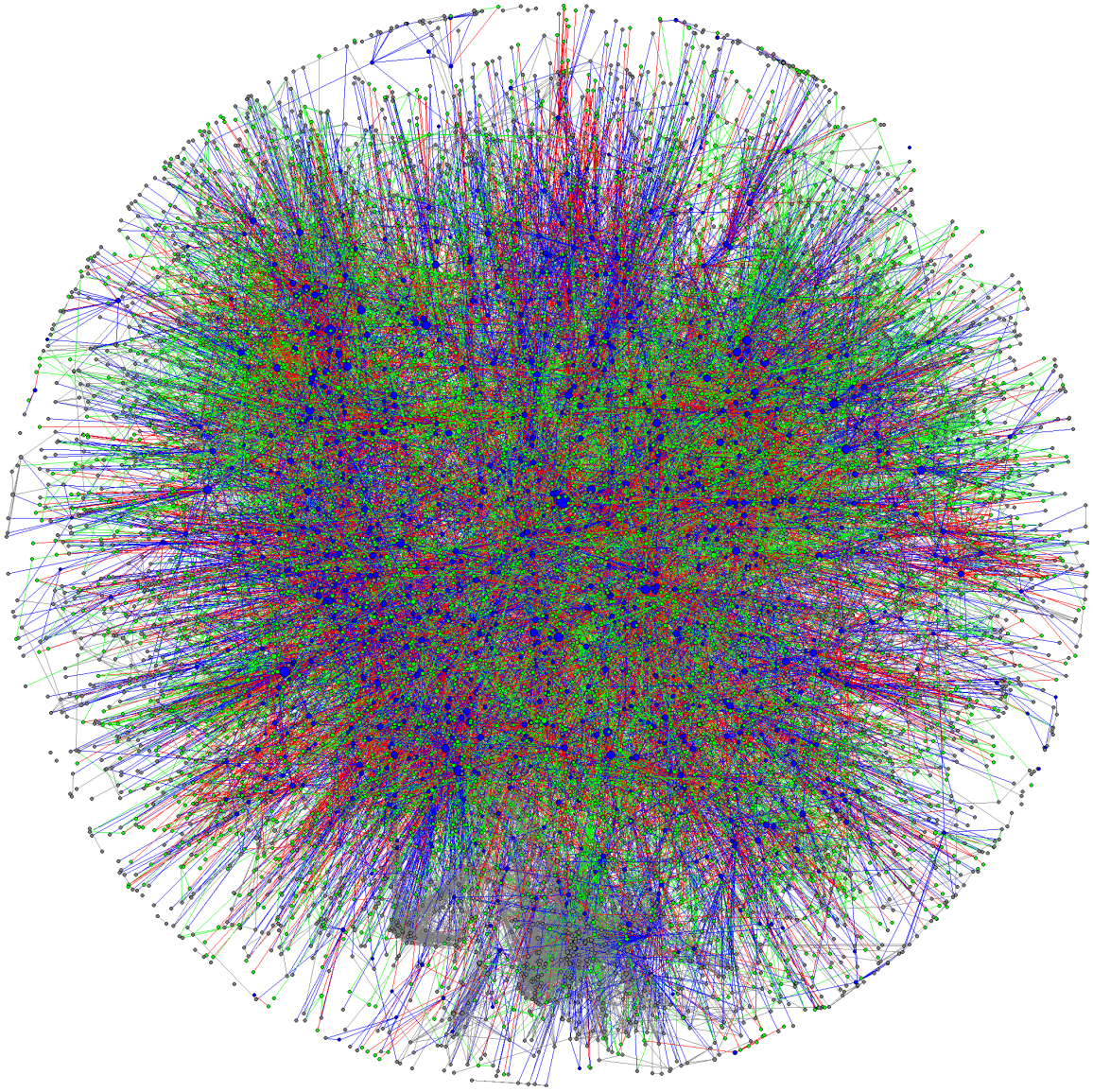
Por fim, foram calculadas as correlações entre as medidas de produtividade e orientações em relação às duas medidas de centralidade extraídas da rede contendo os dados das produções de 2000 a 2012 (isto é, grau e *author rank*). Estas correlações são apresentadas na figura 38. Correlações altas entre as medidas de centralidade e a quantidade de publicações (e, em especial, em relação às publicações mais frequentes) já eram esperadas, pois a rede foi formada a partir de coautorias em publicações. Dois valores altos de correlação que não eram inicialmente esperados ocorreram entre a medida *author rank* e o número de publicações de livros (0,52) e de capítulos de livros (0,58).

7.3.3 Conclusões - relação orientador-orientado

Nesta seção foi analisada, de maneira quantitativa, a participação dos orientados nas publicações de seus orientadores, a partir de dados dos orientadores dos programas de pós-graduação em Ciência da Computação no Brasil.

Observou-se que para este conjunto há intensa participação dos orientados, os quais participam de mais de metade do total das publicações de seus orientadores. Além disto,

Figura 36 – Rede com as coautorias acumuladas de 2000 a 2012



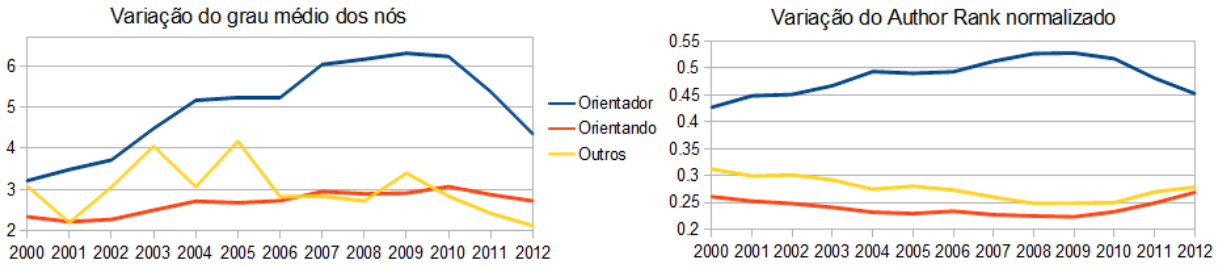
Fonte: Digiampietri, Mugnaini e Alves (2013)

dentre as publicações cujo primeiro autor é o orientador ou o orientado, mais de metade tem um orientado como primeiro autor.

Conforme era esperado, foram identificadas correlações altas entre o número de orientações e o número de publicações e foi possível observar quais tipos de orientações estão mais correlacionadas com quais tipos de publicação.

Também foram analisadas as redes de coautoria envolvendo orientadores, seus orientados e demais colaboradores. Há um indício de que, ao longo dos anos, os orientados têm tido sua importância relativamente aumentada em relação à importância dos demais colaboradores do orientador.

Figura 37 – Variação do grau e do *Author Rank* de 2000 a 2012



Fonte: Digiampietri, Mugnaini e Alves (2013)

Figura 38 – Correlação entre a medida *Author Rank* e grau dos nós e as demais métricas

	total de artigos orientador	artigos completos publicados em periódicos	trabalhos completos publicados em anais de congressos	resumos expandidos publicados em anais de congressos	resumos publicados em anais de congressos	livros publicados organizados ou edições	capítulos de livros publicados	primeira autoria
Author Rank	0,82	0,6	0,8	0,31	0,24	0,52	0,58	-0,44
grau	0,7	0,7	0,66	0,35	0,36	0,22	0,19	-0,39

	total de supervisões	supervisão de pós-doutorado	tese de doutorado	dissertação de mestrado	iniciação científica	trabalho de conclusão de curso	orientação de outra natureza	monografia de conclusão de curso de aperfeiçoamento ou especialização
Author Rank	0,5	0,22	0,63	0,68	0,22	0,11	0,04	-0,02
grau	0,43	0,2	0,59	0,63	0,19	0,08	0,01	-0,04

Fonte: Digiampietri, Mugnaini e Alves (2013)

Apesar de os orientados participarem de mais de metade das publicações dos artigos de seus orientadores e ocuparem mais a função de primeiro autor dos artigos em colaboração, ao se observar medidas de centralidade da rede de coautorias é possível perceber a importância integradora dos orientadores (que ocupam 96 das 100 primeiras posições da rede em relação à medida *author rank*).

8 Conclusões e Trabalhos Futuros

Neste documento foram apresentados diferentes resultados de pesquisas realizadas pelo autor e seus colaboradores sobre a análise da rede social acadêmica brasileira.

Os resultados podem ser agrupados em três tipos. Os resultados metodológicos, que descrevem as diferentes etapas que podem ser utilizadas para a extração, organização, enriquecimento e análise de dados da rede social acadêmica brasileira ou, especificamente, para a utilização de dados oriundos da Plataforma Lattes.

Outros resultados correspondem à caracterização ou detalhamento de redes sociais acadêmicas brasileiras utilizando diferentes medidas bibliométricas, da análise de redes sociais ou oriundas da mineração de textos.

Por fim, há resultados sobre novas estratégias para resolver problemas específicos relacionados à análise de redes sociais, como resolução de entidades, predição de relacionamentos, análise de tendências e identificação de áreas de atuação.

Como trabalhos futuros pretende-se concluir as pesquisas em andamento apresentadas, especialmente a realização de testes e validações adicionais relacionados à predição de relacionamentos e análise de tendências.

Pretende-se também aprimorar a estratégia de desambiguação de nomes de autores e testá-la em conjuntos maiores de dados do projeto DBLP. Esta atividade será utilizada como pré-processamento para um estudo comparativo dos programas de pós-graduação brasileiros em Ciência da Computação considerados de nível internacional com os principais programas internacionais.

Referências[1]

AGGARWAL, C.; SUBBIAN, K. Evolutionary network analysis: A survey. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 47, n. 1, p. 10:1–10:36, maio 2014. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/2601412>. Citado na página 47.

ALABASTRO, P. et al. Mymuseum: Integrating personalized recommendation and multimedia for enriched human-system interaction. In: *Digital Content, Multimedia Technology and its Applications (IDC), 2010 6th International Conference on*. [S.l.: s.n.], 2010. p. 421 –426. Citado na página 46.

ALMIND, T. C.; INGWERSEN, P. Informetric analyses on the world wide web: Methodological approaches to "webometrics". *Journal of Documentation*, v. 53, n. 4, p. 404–426, set. 1997. Citado na página 28.

ALTINTAS, I. et al. Kepler: An extensible system for design and execution of scientific workflows. In: *Proceedings of the 16th International Conference on Scientific and Statistical Database Management*. Washington, DC, USA: [s.n.], 2004. p. 423–424. Citado na página 107.

ALVES, A.; YANASSE, H.; SOMA, N. Sucupira: A system for information extraction of the Lattes platform to identify academic social networks. In: *Information Systems and Technologies (CISTI), 2011 6th Iberian Conference on*. [S.l.: s.n.], 2011. p. 1–6. Citado 2 vezes nas páginas 22 e 30.

ALVES, A. D.; YANASSE, H. H.; SOMA, N. Y. Lattesminer: A multilingual dsl for information extraction from lattes platform. In: *Proceedings of the Compilation of the Co-located Workshops on DSM'11, TMC'11, AGERE! 2011, AOOPES'11, NEAT'11, & VMIL'11*. ACM, 2011. (SPLASH '11 Workshops), p. 85–92. ISBN 978-1-4503-1183-0. Disponível em: <http://doi.acm.org/10.1145/2095050.2095065>. Citado 2 vezes nas páginas 22 e 30.

ALVES, C. M.; DIGIAMPIETRI, L. A. Análise de redes sociais: Desenvolvimento de ferramentas para a análise da comunidade científica Brasileira. In: *Anais do 21 Simpósio Internacional de Iniciação Científica da USP (SIICUSP2013)*. [S.l.: s.n.], 2013. Citado na página 19.

ANWAR, M. A. From doctoral dissertation to publication. A study of 1995 American graduates in library and information science. *Journal of Librarian and Information Science*, v. 36 (4), p. 151–157, 2004. Citado na página 51.

ARRIOLA-QUIROZ, I. et al. Characteristics and publication pattern of theses from a Peruvian medical school. *Health Information and Libraries Journal*, v. 27, p. 148–154, 2010. Citado na página 51.

ARRUDA, D. et al. Brazilian computer science research: Gender and regional distributions. *Scientometrics*, Springer Netherlands, v. 79, n. 3, p. 651–665, 2009. ISSN 0138-9130. Disponível em: <http://dx.doi.org/10.1007/s11192-007-1944-0>. Citado na página 21.

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

BARBIERI, N.; BONCHI, F.; MANCO, G. Who to follow and why: Link prediction with explanations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2014. (KDD '14), p. 1266–1275. ISBN 978-1-4503-2956-9. Disponível em: <http://doi.acm.org/10.1145/2623330.2623733>. Citado na página 48.

BARRAGANS-MARTINEZ, A. et al. Exploiting social tagging in a web 2.0 recommender system. *Internet Computing, IEEE*, v. 14, n. 6, p. 23 –30, 2010. ISSN 1089-7801. Citado na página 46.

BARTAL, A.; SASSON, E.; RAVID, G. Predicting links in social networks using text mining and sna. In: *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*. Washington, DC, USA: IEEE Computer Society, 2009. (ASONAM '09), p. 131–136. ISBN 978-0-7695-3689-7. Disponível em: <http://dx.doi.org/10.1109/ASONAM.2009.12>. Citado na página 50.

BAUMES, J. et al. Visage: A virtual laboratory for simulation and analysis of social group evolution. *ACM Trans. Auton. Adapt. Syst.*, ACM, New York, NY, USA, v. 3, n. 3, p. 8:1–8:35, ago. 2008. ISSN 1556-4665. Disponível em: <http://doi.acm.org/10.1145/1380422.1380423>. Citado na página 48.

BERGER-WOLF, T. Y.; SAIA, J. A framework for analysis of dynamic social networks. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: [s.n.], 2006. p. 523–528. ISBN 1-59593-339-5. Disponível em: <http://doi.acm.org/10.1145/1150402.1150462>. Citado na página 48.

BERKOWITZ, S. D. *An Introduction to Structural Analysis: The Network Approach to Social Research*. [S.l.]: Butterworths, 1982. Citado 2 vezes nas páginas 14 e 21.

BERNARDES, D. et al. A social formalism and survey for recommender systems. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 16, n. 2, p. 20–37, maio 2015. ISSN 1931-0145. Disponível em: <http://doi.acm.org/10.1145/2783702.2783705>. Citado na página 47.

BERRY, M. *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer New York, 2013. ISBN 9781475743050. Disponível em: <https://books.google.com.br-/books?id=alvtBwAAQBAJ>. Citado na página 42.

BOLLEN, J.; RODRIQUEZ, M. A.; SOMPEL, H. Van de. Journal status. *Scientometrics*, Kluwer Academic Publishers, v. 69, p. 669–687, 2006. ISSN 0138-9130. Citado na página 39.

BONACICH, P. Power and centrality: A family of measures. *The American Journal of Sociology*, v. 95, n. 5, p. 1170–1182, 1987. Citado na página 83.

BREIGER, R. The analysis of social networks. In: *Handbook of Data Analysis*. [S.l.]: Sage Publications, 2004. Citado 2 vezes nas páginas 14 e 21.

BRITO, J. F. de; DIGIAMPIETRI, L. A. Uma revisão acerca da recomendação personalizada de conteúdo. *Revista de Sistemas de Informação da FSMA*, v. 12, p. 33–40, 2013. Citado 2 vezes nas páginas 17 e 45.

CALLAHAN, S. P. et al. Managing the evolution of dataflows with vistrails. In: *Proceedings of the 22nd International Conference on Data Engineering Workshops*. Washington, DC, USA: IEEE Computer Society, 2006. (ICDEW '06), p. 71–. ISBN 0-7695-2571-7. Disponível em: <http://dx.doi.org/10.1109/ICDEW.2006.75>. Citado na página 107.

CALLON, M. et al. *Cienciometría: El estudio cuantitativo de la actividad científica: de la bibliometría a la vigilancia tecnológica*. Trea, 1995. ISBN 9788487733949. Disponível em: <https://books.google.com.br/books?id=nTdONAAACAAJ>. Citado 2 vezes nas páginas 14 e 27.

CANIBANO, C.; BOZEMAN, B. Curriculum vitae method in science policy and research evaluation: the state-of-the-art. *Research Evaluation*, v. 18, n. 2, p. 86–94, 2009. Citado 4 vezes nas páginas 14, 32, 33 e 59.

CANTADOR, I.; BELLOGIN, A.; CASTELLS, P. Ontology-based personalised and context-aware recommendations of news items. In: *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on*. [S.l.: s.n.], 2008. v. 1, p. 562 –565. Citado na página 46.

CAZELLA, S. C.; NUNES, M. A. S. N.; REATEGUI, E. Jornada de atualização de informática. In: _____. [S.l.]: SBC, 10. cap. A Ciência do Palpite: Estado da Arte em Sistemas de Recomendação, p. 161–216. Citado na página 44.

CHAGAS, F. M.; PEREZ-ALCAZAR, J. J.; DIGIAMPIETRI, L. A. Algoritmo de classificação de especialistas em áreas na base de currículos Lattes. *Em Questão*, 2015. (aceito para publicação). Citado na página 17.

COHEN, W.; RAVIKUMAR, P.; FIENBERG, S. A comparison of string distance metrics for name-matching tasks. In: *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web*. [S.l.: s.n.], 2003. p. 73–78. Citado na página 60.

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). *Plataforma Lattes*. 2015. http://www.cnpq.br/web/portal-lattes/sobre-a-plataforma. Citado na página 26.

CORMEN, T. H. et al. *Introduction to Algorithms*. Cambridge, MA: MIT Press, 2001. Citado 2 vezes nas páginas 22 e 23.

COSTA, B.; PEDRO, E. da S.; MACEDO, G. de. Scientific collaboration in biotechnology: the case of the northeast region in brazil. *Scientometrics*, Springer Netherlands, v. 95, n. 2, p. 571–592, 2013. ISSN 0138-9130. Disponível em: <http://dx.doi.org/10.1007/s11192-012-0924-1>. Citado na página 21.

CUKIERSKI, W.; HAMNER, B.; YANG, B. Graph-based features for supervised link prediction. In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. [S.l.: s.n.], 2011. p. 1237–1244. ISSN 2161-4393. Citado na página 105.

DHOTE, Y.; MISHRA, N.; SHARMA, S. Survey and analysis of temporal link prediction in online social networks. In: *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*. [S.l.: s.n.], 2013. p. 1178–1183. Citado na página 105.

DIAZ-VALENZUELA, I.; MARTÍN-BAUTISTA, M. J.; VILA, M. A. A fuzzy semisupervised clustering method: Application to the classification of scientific publications. In: A., L. et al. (Ed.). *Information Processing and management of uncertainty in knowledge-based systems - 15th International Conference, IPMU 2014*. [S.l.: s.n.], 2014. p. 179–188. Citado na página 36.

DIESTEL, R. *Graph Theory. Graduate Texts in Mathematics*. 3. ed. New York: Springer-Verlag, 2006. Citado 3 vezes nas páginas 22, 23 e 24.

DIGIAMPIETRI, L. et al. Minerando e Caracterizando Dados de Currículos Lattes. In: *CSBC 2012 - BraSNAM*. [S.l.: s.n.], 2012. Citado 3 vezes nas páginas 54, 55 e 57.

DIGIAMPIETRI, L. et al. Dinâmica das Relações de Coautoria nos Programas de Pós-Graduação em Computação no Brasil. In: *CSBC 2012 - BraSNAM*. [S.l.: s.n.], 2012. Citado 6 vezes nas páginas 41, 48, 62, 63, 76 e 86.

DIGIAMPIETRI, L.; SANTIAGO, C.; ALVES, C. Predição de coautorias em redes sociais acadêmicas: um estudo exploratório em ciência da computação. In: *CSBC-BraSNAM 2013*. [S.l.: s.n.], 2013. Citado 10 vezes nas páginas 48, 102, 103, 105, 109, 110, 111, 112, 113 e 114.

DIGIAMPIETRI, L. et al. Um sistema de informação extensível para o reconhecimento automático de libras. In: *SBSI 2012 - Trilhas Técnicas (Technical Tracks)*. [S.l.: s.n.], 2012. Citado 3 vezes nas páginas 88, 89 e 107.

DIGIAMPIETRI, L. A. et al. Análise da rede de relacionamentos dos doutores brasileiros. In: *VIII Brazilian e-Science Workshop (BRESCI2014) - Anais do XXXIV Congresso da Sociedade Brasileira de Computação (CSBC2014)*. Brasília, DF, Brasil: [s.n.], 2014. p. 323–330. Citado 12 vezes nas páginas 18, 41, 90, 91, 92, 93, 95, 97, 98, 99, 100 e 101.

DIGIAMPIETRI, L. A. et al. Análise da rede dos doutores que atuam em computação no Brasil. In: *III Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2014) - Anais do XXXIV Congresso da Sociedade Brasileira de Computação (CSBC2014)*. [S.l.: s.n.], 2014. p. 33–44. Citado na página 18.

DIGIAMPIETRI, L. A. et al. Combinando workflows e semântica para facilitar o reuso de software. *Revista de Informática Teórica e Aplicada: RITA*, v. 20, p. 73–89, 2013. Citado na página 107.

DIGIAMPIETRI, L. A.; BARBOSA, L. F.; LINDEN, R. Desambiguação de nomes em redes sociais acadêmicas: Um estudo de caso usando DBLP. In: *IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2015) - Anais do XXXV Congresso da Sociedade Brasileira de Computação (CSBC2015)*. [S.l.: s.n.], 2015. Citado 6 vezes nas páginas 19, 69, 70, 71, 72 e 73.

DIGIAMPIETRI, L. A.; MARUYAMA, W. T. Predição de novas coautorias na rede social acadêmica dos programas brasileiros de pós-graduação em ciência da computação. In: *III Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2014) - Anais do XXXIV Congresso da Sociedade Brasileira de Computação (CSBC2014)*. [S.l.: s.n.], 2014. p. 243–248. Citado 2 vezes nas páginas 19 e 102.

DIGIAMPIETRI, L. A. et al. Um sistema de predição de relacionamentos em redes sociais. In: *XI Simpósio Brasileiro de Sistemas de Informação (SBSI 2015)*. [S.l.: s.n.], 2015. p. 139–146. Citado 8 vezes nas páginas 18, 48, 102, 104, 106, 107, 115 e 116.

DIGIAMPIETRI, L. A. et al. Minerando e caracterizando dados de currículos Lattes. In: *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2012) - Anais do XXXII Congresso da Sociedade Brasileira de Computação (CSBC 2012)*. Curitiba, PR, Brasil: [s.n.], 2012. p. 12. Citado na página 18.

DIGIAMPIETRI, L. A. et al. Dinâmica das relações de coautoria nos programas de pós-graduação em computação no Brasil. In: *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2012) - Anais do XXXII Congresso da Sociedade Brasileira de Computação (CSBC 2012)*. Curitiba, PR, Brasil: [s.n.], 2012. p. 12. Citado na página 18.

DIGIAMPIETRI, L. A. et al. BraX-Ray: An X-Ray of the Brazilian Computer Science Graduate Programs. *PLoS ONE*, Public Library of Science, v. 9, n. 4, p. e94541, 04 2014. Disponível em: <http://dx.doi.org/10.1371%2Fjournal.pone.0094541>. Citado 14 vezes nas páginas 17, 38, 41, 48, 75, 77, 78, 79, 80, 81, 84, 85, 86 e 87.

DIGIAMPIETRI, L. A. et al. Extração, caracterização e análises de dados de currículos lattes. *RESI: Revista Eletrônica de Sistemas de Informação*, 2015. (aceito para publicação). Citado na página 17.

DIGIAMPIETRI, L. A. et al. Análise da evolução das relações de coautoria nos programas de pós-graduação em computação no Brasil. *RESI : Revista Eletrônica de Sistemas de Informação*, 2015. (aceito para publicação). Citado 4 vezes nas páginas 17, 41, 48 e 86.

DIGIAMPIETRI, L. A.; MUGNAINI, R.; ALVES, C. M. Análise da participação dos orientandos na produção dos orientadores: um estudo de caso em Ciência da Computação. In: *II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2013) - Anais do XXXIII Congresso da Sociedade Brasileira de Computação (CSBC 2013)*. Maceió, Alagoas, Brasil: [s.n.], 2013. p. 12. Citado 9 vezes nas páginas 18, 52, 134, 135, 136, 137, 138, 139 e 140.

DIGIAMPIETRI, L. A. et al. Análise da atualização dos currículos lattes. In: *Anais do IV Encontro Brasileiro de Bibliometria e Cientometria (EBBC 2014)*. Recife, PE: [s.n.], 2014. p. 8. Citado 2 vezes nas páginas 18 e 59.

DIGIAMPIETRI, L. A. et al. Análise macro das últimas atualizações dos currículos lattes. *Em Questão*, v. 20, p. 88–113, 2014. 2014. Citado 3 vezes nas páginas 17, 59 e 87.

DIGIAMPIETRI, L. A. et al. An extensible framework for genomic and metagenomic analysis. In: CAMPOS, S. (Ed.). *Advances in Bioinformatics and Computational Biology*. Springer International Publishing, 2014, (Lecture Notes in Computer Science, v. 8826). p. 1–8. ISBN 978-3-319-12417-9. Disponível em: <http://dx.doi.org/10.1007/978-3-319-12418-6_1>. Citado na página 107.

DIGIAMPIETRI, L. A.; PERES, S. M.; SILVA, L. A. Rede de relacionamentos brasileira de inteligência artificial e computacional. In: *Anais do XI Encontro Nacional de Inteligência Artificial e Computacional*. [S.l.: s.n.], 2014. p. 6. Citado na página 18.

DIGIAMPIETRI, L. A.; SANTIAGO, C. R. do N.; ALVES, C. M. Predição de coautorias em redes sociais acadêmicas: um estudo exploratório em Ciência da Computação. In: *II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2013) - Anais do XXXIII Congresso da Sociedade Brasileira de Computação (CSBC 2013)*. Maceió, Alagoas, Brasil: [s.n.], 2013. p. 12. Citado na página 18.

DIGIAMPIETRI, L. A.; SILVA, E. E. da. A framework for social network of researchers analysis. *Iberoamerican Journal of Applied Computing*, v. 1, n. 1, p. 1 – 24, 2011. Citado 3 vezes nas páginas 17, 53 e 54.

DIGIAMPIETRI, R. M. ad L. A.; MENA-CHALCO, J. P. Correlation among the scientific production, supervisions and participation in defense examination committees in the brazilian physicists community. In: *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference (ISSI 2013)*. [S.l.: s.n.], 2013. I, p. 447–474. Citado na página 18.

DONG, Y. et al. Random walk based resource allocation: Predicting and recommending links in cross-operator mobile communication networks. In: *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. [S.l.: s.n.], 2011. p. 358–365. Citado na página 49.

DONG, Y. et al. Link prediction and recommendation across heterogeneous social networks. In: *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. [S.l.: s.n.], 2012. p. 181–190. ISSN 1550-4786. Citado na página 48.

DUFFY, R. D. et al. The research productivity of academic psychologists: assessment, trends, and best practice recommendations. *Scientometrics*, v. 89, n. 1, p. 207–227, 2011. Citado na página 39.

EGGHE, L. Theory and practise of the g-index. *Scientometrics*, Springer Netherlands, v. 69, n. 1, p. 131–152, 2006. ISSN 0138-9130. Disponível em: <http://dx.doi.org/10.1007-/s11192-006-0144-7>. Citado na página 29.

FERREIRA, A. A.; GONCALVES, M. A.; LAENDER, A. H. F. A brief survey of automatic methods for author name disambiguation. *SIGMOD Record*, v. 41, n. 2, p. 15–26, 2012. Citado 2 vezes nas páginas 36 e 37.

FERREIRA, A. A.; MACHADO, T. M.; GONCALVES, M. A. Improving author name disambiguation with user relevance feedback. *Journal of Information and Data Management*, v. 3, n. 3, p. 332–347, 2012. Citado na página 37.

FIRE, M. et al. Link prediction in social networks using computationally efficient topological features. In: *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*. [S.l.: s.n.], 2011. p. 73–80. Citado na página 48.

FRANCESCHET, M. A comparison of bibliometric indicators for computer science scholars and journals on web of science and google scholar. *Scientometrics*, Akadémiai Kiadó, co-published with Springer Science+ Business Media BV, Formerly Kluwer Academic Publishers BV, v. 83, n. 1, p. 243–258, 2010. Citado na página 39.

FRANCESCHET, M. Collaboration in computer science: A network science approach. *Journal of the American Society for Information Science and Technology*, Wiley Online Library, v. 62, n. 10, p. 1992–2012, 2011. Citado 2 vezes nas páginas 39 e 40.

FREEMAN, L. C. Centrality in social networks: Conceptual clarification. *Social Networks*, v. 1, p. 215–239, 1979. Citado na página 25.

FREIRE, F.; DIGIAMPIETRI, L. A. Desenvolvimento de sistema de recomendação de artigos científicos. In: *Anais do 19 Simpósio Internacional de Iniciação Científica da USP (SIICUSP2011)*. [S.l.: s.n.], 2011. Citado na página 19.

FREIRE, V.; FIGUEIREDO, D. Ranking in collaboration networks using a group based metric. *Journal of the Brazilian Computer Society*, Springer, p. 1–12, 2011. Citado na página 40.

FRKOVIC, V.; SKENDER, T.; DOJCINOVIC, B. Publishing scientific papers based on master's and Ph.D. theses from a small community: case study Croatian medical schools. *Croatian Medical Journal*, v. 44(1), p. 107–111, 2003. Citado na página 51.

FRUCHTERMAN, T. M. J.; REINGOLD, E. M. Graph drawing by force-directed placement. *Software: Practice and Experience*, John Wiley & Sons, Ltd., v. 21, n. 11, p. 1129–1164, 1991. ISSN 1097-024X. Citado na página 84.

GAO, S.; DENOYER, L.; GALLINARI, P. Link prediction via latent factor blockmodel. In: *Proceedings of the 21st International Conference Companion on World Wide Web*. New York, NY, USA: ACM, 2012. (WWW '12 Companion), p. 507–508. ISBN 978-1-4503-1230-1. Disponível em: <http://doi.acm.org/10.1145/2187980.2188100>. Citado na página 49.

GAO, Y.; XU, B.; CAI, H. Information recommendation method research based on trust network and collaborative filtering. In: *e-Business Engineering (ICEBE), 2011 IEEE 8th International Conference on*. [S.l.: s.n.], 2011. p. 386–391. Citado na página 47.

GARFIELD, E. Citation indexes for science. a new dimension in documentation through association of ideas. *Science*, v. 122, p. 1123–1127, 1955. Citado na página 39.

GERDSRI, N.; KONGTHON, A.; PUENGRUSME, S. Discovering the professional communities and social networks of emerging research areas: Use of technology intelligence from bibliometric and text mining analysis. In: *Technology Management for Emerging Technologies (PICMET 2012)*. [S.l.: s.n.], 2012. p. 114–121. Citado na página 44.

GHAREHCHOPOGH, F. S.; KHALIFELU, Z. A. Analysis and evaluation of unstructured data: text mining versus natural language processing. In: *5th Int. Conf. on Application of Information and Communication Technologies (AICT'2011)*. [S.l.: s.n.], 2011. p. 1–4. Citado na página 44.

GLANZEL, W.; SCHUBERT, A. Analysing scientific networks through coauthorship. In: *Handbook of quantitative science and technology research*. [S.l.]: Kluwer Academic Publishers, 2004. p. 257–276. Citado na página 39.

GLOOR, P. a. et al. Web Science 2.0: Identifying Trends through Semantic Social Network Analysis. *2009 International Conference on Computational Science and Engineering*, Ieee, p. 215–222, 2009. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper-.htm?arnumber=5284145>. Citado na página 42.

GODOI, T. A. et al. A relevance feedback approach for the author name disambiguation problem. In: *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY, USA: ACM, 2013. (JCDL '13), p. 209–218. ISBN 978-1-4503-2077-1. Disponível em: <http://doi.acm.org/10.1145/2467696.2467709>. Citado na página 37.

GOECKS, J. et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, v. 11, n. 8, p. R86, 2010. ISSN 1465-6906. Citado na página 107.

GOLDSCHMIDT, R.; PASSOS, E. *Data mining: um guia Prático*. [S.l.]: CAMPUS, 2005. ISBN 9788535218770. Citado na página 32.

GUO, J.; GUO, H. Multi-features link prediction based on matrix. In: *Computer Design and Applications (ICCDA), 2010 International Conference on*. [S.l.: s.n.], 2010. v. 1, p. V1–357–V1–361. Citado na página 49.

GUO, L. et al. Analyzing patterns of user content generation in online social networks. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2009. (KDD '09), p. 369–378. ISBN 978-1-60558-495-9. Disponível em: <http://doi.acm.org/10.1145/1557019.1557064>. Citado na página 48.

HALL, M. et al. The WEKA data mining software: an update. *SIGKDD Explorations*, v. 11, n. 1, p. 10–18, 2009. Disponível em: <http://doi.acm.org/10.1145/1656274.1656278>. Citado na página 108.

HAN, H. et al. Two supervised learning approaches for name disambiguation in author citations. In: CHEN, H. (Ed.). *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*. [S.l.: s.n.], 2004. p. 296–305. Citado na página 36.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Elsevier Science, 2006. ISBN 9780080475585. Disponível em: <https://books.google.com.br/books?id=AfL0t-YzOrEC>. Citado na página 42.

HASAN, M.; ZAKI, M. A survey of link prediction in social networks. In: AGGARWAL, C. C. (Ed.). *Social Network Data Analytics*. Springer US, 2011. p. 243–275. ISBN 978-1-4419-8461-6. Disponível em: <http://dx.doi.org/10.1007/978-1-4419-8462-3_9>. Citado 2 vezes nas páginas 50 e 105.

HASAN, M. A. et al. Link prediction using supervised learning. In: *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*. [S.l.: s.n.], 2006. Citado na página 105.

HAYAT, Z.; LYONS, K. The evolution of the cascon community: a social network analysis. In: *Proceedings of the 2010 Conference of the Center for Advanced Studies on Collaborative Research*. Riverton, NJ, USA: IBM Corp., 2010. (CASCON '10), p. 1–12. Disponível em: <http://dx.doi.org/10.1145/1923947.1923949>. Citado na página 48.

HE, H. et al. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*. [S.l.: s.n.], 2008. p. 1322–1328. ISSN 1098-7576. Citado na página 106.

HIRSCH, J. E. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, v. 102, n. 46, p. 16569–16572, 2005. Citado 2 vezes nas páginas 29 e 39.

HORN, D. B. et al. Six degrees of jonathan grudin: a social network analysis of the evolution and impact of cscw research. In: *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. New York, NY, USA: ACM, 2004. (CSCW '04), p. 582–591. ISBN 1-58113-810-5. Disponível em: <http://doi.acm.org/10.1145/1031607.1031707>. Citado na página 48.

HOSEINI, E.; HASHEMI, S.; HAMZEH, A. Link prediction in social network using co-clustering based approach. In: *26th International Conference on Advanced Information Networking and Applications Workshops*. [S.l.: s.n.], 2012. p. 795–800. Citado na página 50.

HSIEH, C.-J. et al. Organizational overlap on social networks and its applications. In: *Proceedings of the 22Nd International Conference on World Wide Web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. (WWW '13), p. 571–582. ISBN 978-1-4503-2035-1. Disponível em: <http://dl.acm.org/citation.cfm?id=2488388.2488439>. Citado na página 48.

IAQUINTA, L. et al. Recommendations toward serendipitous diversions. In: *Intelligent Systems Design and Applications, 2009. ISDA '09. Ninth International Conference on*. [S.l.: s.n.], 2009. p. 1049 –1054. Citado na página 46.

IBRAHIM, N.; CHEN, L. Link prediction in dynamic social networks by integrating different types of information. *Applied Intelligence*, Springer US, v. 42, n. 4, p. 738–750, 2015. ISSN 0924-669X. Disponível em: <http://dx.doi.org/10.1007/s10489-014-0631-0>. Citado na página 50.

IGAMI, M. P. Z.; BRESSIANI, J. C.; MUGNAINI, R. A new model to identify the productivity of theses in terms of articles using co-word analysis. *Journal of Scientometric Research*, v. 3, p. 3–14, 2014. Citado na página 51.

JAMALI, M.; ESTER, M. Trustwalker: A random walk model for combining trust-based and item-based recommendation. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2009. (KDD '09), p. 397–406. ISBN 978-1-60558-495-9. Disponível em: <http://doi.acm.org/10.1145/1557019.1557067>. Citado na página 47.

JAMALI, M.; ESTER, M. A matrix factorization technique with trust propagation for recommendation in social networks. In: *Proceedings of the Fourth ACM Conference on Recommender Systems*. New York, NY, USA: ACM, 2010. (RecSys '10), p. 135–142. ISBN 978-1-60558-906-0. Disponível em: <http://doi.acm.org/10.1145/1864708.1864736>. Citado na página 47.

JOLLIFFE, I. T. *Principal Component Analysis*. Second. [S.l.]: Springer, 2002. Citado 2 vezes nas páginas 84 e 85.

JULASHOKRI, M.; FATHIAN, M.; GHOLAMIAN, M. Improving customer's profile in recommender systems using time context and group preferences. In: *Computer Sciences and Convergence Information Technology (ICCIT), 2010 5th International Conference on*. [S.l.: s.n.], 2010. p. 125 –129. Citado na página 46.

KANG, H.; GETOOR, L.; SINGH, L. Visual analysis of dynamic group membership in temporal social networks. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 9, n. 2, p. 13–21, dez. 2007. ISSN 1931-0145. Disponível em: <http://doi.acm.org/10.1145-/1345448.1345452>. Citado na página 48.

KANO, V. Y.; MIYATA, B. K. O.; DIGIAMPIETRI, L. A. Uso de mineração de textos para análise de características da produção científica nacional. In: *Anais do 21 Simpósio Internacional de Iniciação Científica da USP (SIICUSP2013)*. [S.l.: s.n.], 2013. Citado na página 19.

KONTOSTATHIS, A.; GALITSKY, L.; POTTENGER, W. A survey of emerging trend detection in textual data mining. *Survey of Text ...*, p. 1–44, 2004. Disponível em: <http://link.springer.com/chapter/10.1007/978-1-4757-4305-0_9>. Citado na página 42.

LAENDER, A. et al. Assessing the research and education quality of the top Brazilian Computer Science graduate programs. *ACM Special Interest Group on Computer Science Education*, ACM, v. 40, n. 2, p. 135–145, 2008. Citado 2 vezes nas páginas 38 e 40.

LAENDER, A. H. et al. Building a research social network from an individual perspective. In: *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. New York, NY, USA: ACM, 2011. (JCDL '11), p. 427–428. ISBN 978-1-4503-0744-4. Disponível em: <http://doi.acm.org/10.1145/1998076.1998168>. Citado na página 30.

LANGVILLE, A.; MEYER, C. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. [S.l.]: Princeton University Press, 2009. ISBN 9780691122021. Citado na página 83.

LARIVIÈRE, V. On the shoulders of students? the contribution of Phd students to the advancement of knowledge. *Scientometrics*, v. 90, n. 2, p. 463–481, 2012. Citado na página 51.

LEE, W. M. Publication trends of doctoral students in three fields from 1965-1995. *Journal of the American Society for Information Science*, v. 51(2), p. 139–144, 2000. Citado na página 51.

LEITE, P.; MUGNAINI, R.; LETA, J. A new indicator for international visibility: exploring brazilian scientific community. *Scientometrics*, v. 88, p. 311–319, 2011. Citado na página 21.

LEMIEUX, V.; OUIMET, M. *Análise Estrutural das Redes Sociais*. [S.l.]: Instituto Piaget, 2008. 128 p. ISBN 9727719333. Citado 6 vezes nas páginas 14, 21, 22, 23, 25 e 94.

LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, v. 10, n. 8, p. 707–710, 1966. Citado 2 vezes nas páginas 61 e 63.

LEY, M. The dblp computer science bibliography: Evolution, research issues, perspectives. In: *Proceedings of the 9th International Symposium on String Processing and Information Retrieval*. London, UK, UK: Springer-Verlag, 2002. (SPIRE 2002), p. 1–10. ISBN 3-540-44158-1. Citado na página 40.

LIBEN-NOWELL, D.; KLEINBERG, J. The link prediction problem for social networks. In: *Proceedings of the Twelfth International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2003. (CIKM '03), p. 556–559. ISBN 1-58113-723-0. Disponível em: <http://doi.acm.org/10.1145/956863.956972>. Citado 2 vezes nas páginas 48 e 105.

LIMA, J. J. da S.; DIGIAMPIETRI, L. A. Enriquecendo bases de dados de currículos lattes. In: *Anais do 21 Simpósio Internacional de Iniciação Científica da USP (SIICUSP2013)*. [S.l.: s.n.], 2013. Citado na página 19.

LIMA, J. J. da S.; DIGIAMPIETRI, L. A. Iniciação científica e tecnológica: O jovem pesquisador em ação iv. In: _____. [S.l.]: CETEPE/EESC/USP, 2014. cap. Enriquecendo base de dados de currículos Lattes, p. 281–295. Citado na página 19.

LIN, Z.; YUN, X.; ZHU, Y. Link prediction using benefitranks in weighted networks. In: *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*. Washington, DC, USA: IEEE Computer Society, 2012. (WI-IAT '12), p. 423–430. ISBN 978-0-7695-4880-7. Disponível em: <http://dl.acm.org/citation.cfm?id=2457524.2457624>. Citado na página 49.

LIU, X. et al. Co-authorship networks in the digital library research community. *Inf. Process. Manage.*, v. 41, n. 6, p. 1462–1480, 2005. Citado 3 vezes nas páginas 41, 86 e 137.

LOH, S. et al. Comparing keywords and taxonomies in the representation of users profiles in a content-based recommender system. In: *Proceedings of the 2008 ACM symposium on Applied computing*. ACM, 2008. (SAC '08), p. 2030–2034. ISBN 978-1-59593-753-7. Disponível em: <http://doi.acm.org/10.1145/1363686.1364177>. Citado na página 46.

LONG, J.; MCGINNIS, R. The effects of the mentor on the academic career. *Scientometrics*, Kluwer Academic Publishers, v. 7, n. 3-6, p. 255–280, 1985. ISSN 0138-9130. Disponível em: <http://dx.doi.org/10.1007/BF02017149>. Citado na página 51.

LOPS, P. et al. Content-based filtering with tags: The first system. In: *Intelligent Systems Design and Applications, 2009. ISDA '09. Ninth International Conference on*. [S.l.: s.n.], 2009. p. 255 –260. Citado na página 46.

LOVINS, J. B. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, v. 11, n. 1-2, p. 22–31, 1968. Citado na página 118.

LU, L. et al. Recommender systems. *Physics Reports*, v. 519, n. 1, p. 1 – 49, 2012. ISSN 0370-1573. Recommender Systems. Disponível em: <http://www-.sciencedirect.com/science/article/pii/S0370157312000828>. Citado na página 44.

LU, L.; ZHOU, T. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, v. 390, n. 6, p. 1150 – 1170, 2011. ISSN 0378-4371. Disponível em: <http://www.sciencedirect.com/science/article/pii/S037843711000991X>. Citado 2 vezes nas páginas 48 e 105.

MA, H. et al. Recommender systems with social regularization. In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. New York, NY,

USA: ACM, 2011. (WSDM '11), p. 287–296. ISBN 978-1-4503-0493-1. Disponível em: <http://doi.acm.org/10.1145/1935826.1935877>. Citado na página 47.

MAKREHCHI, M. Social link recommendation by learning hidden topics. In: *Proceedings of the Fifth ACM Conference on Recommender Systems*. New York, NY, USA: ACM, 2011. (RecSys '11), p. 189–196. ISBN 978-1-4503-0683-6. Disponível em: <http://doi.acm.org/10.1145/2043932.2043968>. Citado na página 49.

MALLETTE, L. Publishing rates of graduates education Ph.D. and Ed.D. students: a longitudinal study of University of California schools. *Doctoral dissertation, Pepperdine University. Retrieved from ProQuest Dissertations & Theses database. (UMI No. 3239922)*, 2006. Citado na página 51.

MARQUES, K. C. A plataforma Lattes e a organização da informação. *Gestão & Planejamento*, v. 11, p. 250–266, 2010. Citado na página 33.

MARTINS, W. S. et al. Assessing the quality of scientific conferences based on bibliographic citations. *Scientometrics*, v. 83, n. 1, p. 133–155, 2010. Citado na página 39.

MEDEIROS, C. et al. WOODSS and the Web: Annotating and Reusing Scientific Workflows. *ACM SIGMOD Record*, v. 34, n. 3, p. 18–23, 2005. Citado na página 107.

MEDEIROS, C. B.; MENA-CHALCO, J. P. The dynamics of multidisciplinary research networks - mining a public repository of scientists CVs. In: COUNCIL, I. S. S. (Ed.). *World Social Science Forum 2013*. Montreal, Canada: [s.n.], 2013. p. 1–17. Citado na página 41.

MELO-MINARDI, R. et al. Caracterização dos programas de pós-graduação em bioinformática no Brasil. In: *II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2013) - Anais do XXXIII Congresso da Sociedade Brasileira de Computação (CSBC 2013)*. Maceió, Alagoas, Brasil: [s.n.], 2013. p. 12. Citado na página 18.

MELO, P. L. da Cunha e. *Produtividade, Internacionalização e Visibilidade da Comunidade Científica Brasileira na Virada do Milênio*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2011. Citado na página 41.

MELO, P. O. S. Vaz de; ALMEIDA, V. A. F.; LOUREIRO, A. A. F. Can complex network metrics predict the behavior of nba teams? In: *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2008. p. 695–703. ISBN 978-1-60558-193-4. Citado na página 83.

MENA-CHALCO, J.; DIGIAMPIETRI, L.; CESAR-JUNIOR, R. Caracterizando as redes de coautoria de currículos Lattes. In: *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2012) - Anais do XXXII Congresso da Sociedade Brasileira de Computação (CSBC 2012)*. Curitiba, PR, Brasil: [s.n.], 2012. p. 12. Citado na página 18.

MENA-CHALCO, J. P.; CESAR-JUNIOR, R. M. ScriptLattes: an open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, scielo, v. 15, p. 31 – 39, 12 2009. ISSN 0104-6500. Disponível em: <http://www.scielo.br-/scielo.php?script=sci_arttext&pid=S0104-65002009000400004&nrm=iso>. Citado 5 vezes nas páginas 22, 29, 39, 61 e 64.

MENA-CHALCO, J. P.; CESAR-JUNIOR., R. M. Towards automatic discovery of co-authorship networks in the brazilian academic areas. In: *IEEE Seventh International Conference on e-Science Workshops 2011 (eScienceW)*. [S.l.]: IEEE, 2011. p. 53–60. Citado na página 41.

MENA-CHALCO, J. P. et al. Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, v. 65, p. 1424–1445, 2014. ISSN 2330-1643. Disponível em: <http://dx.doi.org/10.1002/asi.23010>. Citado 4 vezes nas páginas 17, 41, 48 e 49.

MENA-CHALCO, J. P.; DIGIAMPIETRI, L. A.; OLIVEIRA, L. B. Perfil de produção acadêmica dos programas brasileiros de pós-graduação em ciencia da computação nos triênios 2004-2006 e 2007-2009. *Em Questão*, v. 18, p. 215–229, 2012. ISSN 1807-8893. Citado 2 vezes nas páginas 17 e 18.

MENEZES, G. V. et al. A geographical analysis of knowledge production. In: *in Computer Science In Proceedings of the 18th international conference on World Wide Web*. [S.l.: s.n.], 2009. p. 1041–1050. Citado 2 vezes nas páginas 39 e 40.

MEYFFRET, S.; MÉDINI, L.; LAFOREST, F. Trust-based local and social recommendation. In: *Proceedings of the 4th ACM RecSys Workshop on Recommender Systems and the Social Web*. New York, NY, USA: ACM, 2012. (RSWeb '12), p. 53–60. ISBN 978-1-4503-1638-5. Disponível em: <http://doi.acm.org/10.1145/2365934.2365945>. Citado na página 47.

MILOJEVIC, S. Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics*, v. 7, n. 2, p. 767–773, 2013. Citado na página 35.

MIYATA, B. K. O.; KANO, V. Y.; DIGIAMPIETRI, L. A. Combinando mineração de textos e análise de redes sociais para a identificação das áreas de atuação de pesquisadores. In: *II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2013) - Anais do XXXIII Congresso da Sociedade Brasileira de Computação (CSBC 2013)*. Maceió, Alagoas, Brasil: [s.n.], 2013. p. 12. Citado 8 vezes nas páginas 18, 117, 119, 120, 121, 122, 123 e 124.

MIYATA, B. K. O.; KANO, V. Y.; DIGIAMPIETRI, L. A. Uso de mineração de textos para a identificação das áreas de atuação de pesquisadores. In: *Anais do 21 Simpósio Internacional de Iniciação Científica da USP (SIICUSP2013)*. [S.l.: s.n.], 2013. Citado na página 19.

MUGNAINI, R.; DIGIAMPIETRI, L. A. The brazilian national impact: movement of journals between Bradford Zones of production and consumption. In: *Proceedings of the 15th International Conference on Scientometrics and Informetrics (ISSI 2015)*. [S.l.: s.n.], 2015. p. 19–24. Citado na página 18.

MUGNAINI, R.; DIGIAMPIETRI, L. A.; MENA-CHALCO, J. P. Comunicação científica no brasil (1998-2012): indexação, crescimento, fluxo e dispersão. *Transinformação*, v. 26, p. 239–252, 2014. Citado na página 17.

MUGNAINI, R.; DIGIAMPIETRI, L. A.; MENA-CHALCO, J. P. Comunicação científica no brasil (1998-2012): infraestrutura nacional e internacionalização. In: *Memórias del XIII*

Congreso Internacional de Información (INFO'2014). Havana, Cuba: [s.n.], 2014. Citado na página 18.

MUGNAINI, R. et al. Normalização de nomes de autores em fontes de informação institucionais: proposta de um método automático de verificação de erros. *Em Questão*, v. 18, n. 3, p. 263–279, 2012. Citado 4 vezes nas páginas 17, 65, 66 e 67.

MUGNAINI, R.; IGAMI, M. P. Z.; BRESSIANI, J. C. Productivity and doctoral research in a brazilian nuclear research institution: validating co-word analysis technique. *Proceedings of the ISSI 2011 Conference. Durban : University of Zululand Reprographic and Printing Centre, 2011.*, v.II, p. 1037–1039, 2011. Citado na página 51.

MUGNAINI, R.; LEITE, P.; LETA, J. Fontes de informação para análise de internacionalização da produção científica brasileira. *PontodeAcesso*, v. 5, n. 3, 2011. ISSN ISSN 1981-6766. Disponível em: <http://www.portalseer.ufba.br/index.php/revistaici-/article/view/5684>. Citado na página 18.

MURATA, T.; MORIYASU, S. Link prediction based on structural properties of online social networks. *New Generation Computing*, Verlag Omsha Tokio, v. 26, n. 3, p. 245–257, 2008. ISSN 0288-3635. Disponível em: <http://dx.doi.org/10.1007/s00354-008-0043-y>. Citado na página 105.

NAKAGAWA, H.; MORI, T. A Simple but Powerful Automatic Term Extraction Method. In: *COLING-02 on COMPUTERM 2002: Second International Workshop on Computational Terminology - Volume 14*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (COMPUTERM '02), p. 1–7. Disponível em: <http://dx.doi.org/10.3115/1118771.1118778>. Citado na página 125.

NARAYANAN, A.; SHI, E.; RUBINSTEIN, B. Link prediction by de-anonymization: How we won the kaggle social network challenge. In: *International Joint Conference on Neural Networks*. [S.l.: s.n.], 2011. p. 1825–1834. ISSN 2161-4393. Citado na página 50.

NEWMAN, M. E. J. Mixing patterns in networks. *Physical Review E*, American Physical Society, v. 67, n. 2, p. 026126, 2003. Citado na página 96.

NEWMAN, M. E. J. The structure and function of complex networks. *SIAM Review*, v. 45, n. 2, p. 167–256, 2003. Disponível em: <http://dx.doi.org/10.1137/S003614450342480>. Citado 2 vezes nas páginas 22 e 24.

NEWMAN, M. E. J. Coauthorship networks and patterns of scientific collaboration. In: *National Academy of Sciences*. [S.l.: s.n.], 2004. p. 5200–5205. Citado 2 vezes nas páginas 39 e 82.

OKAZAKI, N.; TSUJII, J. Simple and efficient algorithm for approximate dictionary matching. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 23rd International Conference on Computational Linguistics*. [S.l.], 2010. p. 851–859. Citado na página 60.

OTTE, E.; ROUSSEAU, R. Social network analysis: a powerful strategy, also for the information sciences. *J. Information Science*, v. 28, n. 6, p. 441–453, 2002. Disponível em: <http://dx.doi.org/10.1177/016555150202800601>. Citado na página 25.

PARK, D. H. et al. A literature review and classification of recommender systems research. *Expert Systems with Applications*, v. 39, n. 11, p. 10059 – 10072, 2012. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417412002825>. Citado na página 45.

PEREZ-ALCAZAR, J. et al. Avaliação de redes de inovação usando uma ferramenta baseada em redes sociais - caso brasileiro de nanotecnologia. In: *Anais do XIV Congreso Latino-Iberoamericano de Gestión Tecnológica (ALTEC 2011)*. [S.l.: s.n.], 2011. Citado na página 18.

PEREZ, C.; BIRREGAH, B.; LEMERCIER, M. The multi-layer imbrication for data leakage prevention from mobile devices. In: *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on*. [S.l.: s.n.], 2012. p. 813–819. Citado na página 48.

PINHEIRO, D.; MELKERS, J.; YOUTIE, J. Learning to play the game: Student publishing as an indicator of future scholarly success. *Technological Forecasting & Social Change*, v. 81, p. 56–66, 2014. Citado na página 51.

POBLACION, D.; MUGNAINI, R.; RAMOS, L. *Redes sociais e colaborativas em informacao cientifica*. 1st. ed. [S.l.]: Angellara Editoras, Sao Paulo, 2009. Citado 2 vezes nas páginas 14 e 22.

PRELL, C. *Social network analysis history, theory & methodology*. [S.l.]: Los Angeles London SAGE, 2012. 263 p. Citado 2 vezes nas páginas 14 e 22.

PRITCHARD, A. Statistical bibliography or bibliometrics? *Journal of Documentation*, v. 25, n. 4, p. 348–349, 1969. Citado 2 vezes nas páginas 14 e 27.

PUDHIYAVEETIL, A. K. et al. Conceptual recommender system for citeseerx. In: *Proceedings of the third ACM conference on Recommender systems*. ACM, 2009. (RecSys '09), p. 241–244. ISBN 978-1-60558-435-5. Disponível em: <http://doi.acm.org/10.1145/1639714.1639758>. Citado na página 46.

QUERCIA, D.; CAPRA, L. Friendsensing: Recommending friends using mobile phones. In: *Proceedings of the Third ACM Conference on Recommender Systems*. New York, NY, USA: ACM, 2009. (RecSys '09), p. 273–276. ISBN 978-1-60558-435-5. Disponível em: <http://doi.acm.org/10.1145/1639714.1639766>. Citado na página 48.

RAMOS, P. et al. Dissertações e Teses de Pós Graduação geram publicação de artigos científicos? análise baseada em 3 programas da área de educação física. *Brazilian Journal Biomotricity*, v. 3(4), p. 315–324, 2009. Citado na página 51.

RATTANAJITBANJONG, N.; MANEEROJ, S. Multi criteria pseudo rating and multidimensional user profile for movie recommender system. In: *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*. [S.l.: s.n.], 2009. p. 596 –601. Citado na página 46.

REATEGUI, E. B.; CAZELLA, S. C. *Sistemas de recomendaçao*. [S.l.], 2005. Citado 2 vezes nas páginas 44 e 45.

RESNICK, P.; VARIAN, H. R. Recommender systems. *Communications of ACM*, ACM, New York, NY, USA, v. 40, n. 3, p. 56–58, mar. 1997. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/245108.245121>. Citado na página 44.

RICCI, F.; ROKACH, L.; SHAPIRA, B. *Introduction to Recommender Systems Handbook*. Springer US, 2011. 1-35 p. ISBN 978-0-387-85819-7. Disponível em: <http://dx.doi.org/10.1007/978-0-387-85820-3_1>. Citado na página 47.

RICCI, F. et al. (Ed.). *Recommender Systems Handbook*. [S.l.]: Springer, 2011. ISBN 978-0-387-85819-7. Citado na página 44.

RODRIGUEZ, J.; KUNCHEVA, L.; ALONSO, C. Rotation forest: A new classifier ensemble method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 28, n. 10, p. 1619–1630, Oct 2006. ISSN 0162-8828. Citado na página 73.

ROWLEY, J. The controlled versus natural indexing languages debate revisited: A perspective on information retrieval practice and research. *J. Inf. Sci.*, Sage Publications, Inc., Thousand Oaks, CA, USA, v. 20, n. 2, p. 108–119, fev. 1994. ISSN 0165-5515. Disponível em: <http://dx.doi.org/10.1177/016555159402000204>. Citado na página 43.

SA, H. de; PRUDENCIO, R. Supervised link prediction in weighted networks. In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. [S.l.: s.n.], 2011. p. 2281–2288. ISSN 2161-4393. Citado na página 48.

SACARDO, M.; HAYASHI, M. C. Balanço bibliométrico da produção científica em Educação Física e Educação Especial oriunda de teses e dissertações. *RBPG Revista Brasileira de Pós-Graduação*, v. 8(15), p. 111–135, 2011. Citado na página 51.

SALMI, L.; GANA, S.; MOUILLET, E. Publication pattern of medical theses, France 1993-98. *Medical Education*, v. 35(1), p. 18–21, 2001. Citado na página 51.

SANTIAGO, C. R. do N. *Desenvolvimento de um ambiente de computação voluntária baseado em computação ponto-a-ponto*. Dissertação (Mestrado) — Universidade de Sao Paulo, 2015. Citado na página 107.

SCOTT, J. *Social network analysis: a handbook*. 2. ed. [S.l.]: SAGE, 2009. Citado 5 vezes nas páginas 14, 21, 22, 24 e 25.

SHARMA, M.; URS, S. R. Network dynamics of scholarship: a social network analysis of digital library community. In: *Proceedings of the 2nd PhD workshop on Information and knowledge management*. New York, NY, USA: [s.n.], 2008. p. 101–104. ISBN 978-1-60558-257-3. Disponível em: <http://doi.acm.org/10.1145/1458550.1458570>. Citado na página 48.

SILVA, F. M.; SMIT, J. W. Organização da informação em sistemas eletrônicos abertos de Informação Científica & Tecnológica: análise da Plataforma Lattes. *Perspectivas em Ciência da Informação*, scielo, v. 14, p. 77 – 98, 04 2009. ISSN 1413-9936. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362009000100007&nrm=iso>. Citado na página 33.

SILVA, G. S.; DIGIAMPIETRI, L. A. Análise de redes sociais de pesquisadores baseada em dados da plataforma Lattes. In: *Anais do 20 Simpósio Internacional de Iniciação Científica da USP (SIICUSP2012)*. [S.l.: s.n.], 2012. Citado na página 19.

SMALHEISER, N. R.; TORVIK, V. I. Author name disambiguation. *Annual Review of Information Science and Technology*, Wiley Subscription Services, Inc., A Wiley Company, v. 43, n. 1, p. 1–43, 2009. ISSN 1550-8382. Disponível em: <http://dx.doi.org/10.1002/aris.2009.1440430113>. Citado na página 35.

SONG, W. et al. Question similarity calculation for faq answering. In: *SKG '07: Proceedings of the Third International Conference on Semantics, Knowledge and Grid*. Washington, DC, USA: IEEE Computer Society, 2007. p. 298–301. ISBN 0-7695-3007-9. Citado na página 36.

SPINAK, E. Indicadores cienciometricos. *Ciência da Informação*, v. 27, n. 2, 1998. ISSN 1518-8353. Disponível em: <http://revista.ibict.br/cienciadainformacao/index.php/ciinf-/article/view/349>. Citado 2 vezes nas páginas 14 e 28.

STROTMANN, A.; ZHAO, D. Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, v. 63, n. 9, p. 1820–1833, 2012. Citado na página 35.

SUN, Y. et al. Co-author relationship prediction in heterogeneous bibliographic networks. In: *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. [S.l.: s.n.], 2011. p. 121–128. Citado na página 50.

SUN, Y. et al. When will it happen?: relationship prediction in heterogeneous information networks. In: *Proceedings of the fifth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, 2012. (WSDM '12), p. 663–672. ISBN 978-1-4503-0747-5. Disponível em: <http://doi.acm.org/10.1145/2124295.2124373>. Citado na página 50.

SZWARCFITER, J. L. *Grafos e algoritmos computacionais*. [S.l.]: Campus, 1986. 216 p. Citado 2 vezes nas páginas 22 e 24.

TAGUE-SUTCLIFFE, J. An introduction to informetrics. *Information Processing & Management*, v. 28, n. 1, p. 1 – 3, 1992. ISSN 0306-4573. Disponível em: <http://www-.sciencedirect.com/science/article/pii/030645739290087G>. Citado 3 vezes nas páginas 14, 27 e 28.

TALBURT, J. R. *Entity Resolution and Information Quality*. 1st. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2010. ISBN 0123819725, 9780123819727. Citado 2 vezes nas páginas 33 e 34.

TAYLOR, I. et al. Visual grid workflow in triana. *Journal of Grid Computing*, v. 3, n. 3-4, p. 153–169, 2005. Citado na página 107.

TIAN, Y. et al. Boosting social network connectivity with link revival. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2010. (CIKM '10), p. 589–598. ISBN 978-1-4503-0099-5. Disponível em: <http://doi.acm.org/10.1145/1871437.1871514>. Citado 2 vezes nas páginas 48 e 105.

TRUCOLO, C. C. *Análise de Tendências em Redes Sociais Acadêmicas*. Dissertação (Mestrado) — Universidade de Sao Paulo, 2015. Citado 3 vezes nas páginas 125, 128 e 132.

TRUCOLO, C. C.; DIGIAMPIETRI, L. A. Análise de tendências da produção científica nacional da área de Ciência da Computação. *Revista de Sistemas de Informação da FSMA*, v. 14, p. 2–10, 2014. 2014. Citado 7 vezes nas páginas 17, 126, 127, 128, 129, 130 e 131.

TRUCOLO, C. C.; DIGIAMPIETRI, L. A. Uma revisão sistemática acerca das técnicas de identificação e análise de tendências. In: *Anais X Simpósio Brasileiro de Sistemas de Informação (SBSI 2014)*. [S.l.: s.n.], 2014. p. 639–650. Citado 2 vezes nas páginas 18 e 43.

TUESTA, E. F. et al. Análise temporal da relação orientador-orientado: um estudo de caso sobre a produtividade dos pesquisadores doutores da área de ciência da computação. In: *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2012) - Anais do XXXII Congresso da Sociedade Brasileira de Computação (CSBC 2012)*. Curitiba, PR, Brasil: [s.n.], 2012. p. 12. Citado 2 vezes nas páginas 18 e 52.

TUESTA, E. F. et al. Análise comparativa da produtividade dos pares orientador-orientado em ciência da computação. *RESI: Revista Eletrônica de Sistemas de Informação*, 2015. (aceito para publicação). Citado 2 vezes nas páginas 17 e 52.

TUESTA, E. F. et al. Analysis of an advisor-advisee relationship: An exploratory study of the area of exact and earth sciences in brazil. *Plos One*, v. 10, p. e0129065, 2015. Citado 2 vezes nas páginas 17 e 52.

ULRIK, B.; ERLEBACH, T. *Network Analysis: Methodological Foundations*. [S.l.]: Springer-Verlag, 2005. Citado 2 vezes nas páginas 14 e 21.

VANTI, N. A. P. Da bibliometria á webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. v. 31, n. 2, p. 152–162, 2002. Citado na página 28.

VARDI, M. Y. Conferences vs. journals in computing research. *Communications of the ACM*, v. 52, n. 5, p. 5, 2009. Citado na página 39.

VASUKI, V. et al. Affiliation recommendation using auxiliary networks. In: *Proceedings of the Fourth ACM Conference on Recommender Systems*. New York, NY, USA: ACM, 2010. (RecSys '10), p. 103–110. ISBN 978-1-60558-906-0. Disponível em: <http://doi.acm.org/10.1145/1864708.1864731>. Citado na página 48.

VEJLGAARD, H. *Anatomy of a Trend*. [S.l.]: McGrawHill, 2008. ISBN 9780071488709. Citado na página 42.

WAINER, J.; VIEIRA, P. Correlations between bibliometrics and peer evaluation for all disciplines: the evaluation of brazilian scientists. *Scientometrics*, Springer Netherlands, v. 96, n. 2, p. 395–410, 2013. ISSN 0138-9130. Disponível em: <http://dx.doi.org/10.1007/s11192-013-0969-9>. Citado na página 21.

WAN, X. et al. Applying keyword map based learner profile to a recommender system for group learning support. In: *Education Technology and Computer Science (ETCS), 2010 Second International Workshop on*. [S.l.: s.n.], 2010. v. 1, p. 3 –6. Citado na página 46.

WANG, C. et al. Mining advisor-advisee relationships from research publication networks. In: *Knowledge Data Discovery (KDD)*. [S.l.: s.n.], 2010. p. 203–212. Citado na página 52.

WANG, C. et al. Learning hierarchical relationships among partially ordered objects with heterogeneous attributes and links. In: *Proceedings of 2012 SIAM International Conference on Data Mining*. [S.l.: s.n.], 2012. p. 516–527. Citado na página 52.

WANG, P. et al. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, Science China Press, v. 58, n. 1, p. 1–38, 2015. ISSN 1674-733X. Disponível em: <http://dx.doi.org/10.1007/s11432-014-5237-y>. Citado na página 50.

WANG, T.; KRIM, H. Statistical classification of social networks. In: *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. [S.l.: s.n.], 2012. p. 3977–3980. ISSN 1520-6149. Citado na página 44.

WANG, T.; KRIM, H.; VINIOTIS, Y. A generalized markov graph model: Application to social network analysis. *Selected Topics in Signal Processing, IEEE Journal of*, v. 7, n. 2, p. 318–332, 2013. ISSN 1932-4553. Citado na página 44.

WASSERMAN, S.; FAUST, K. *Social network analysis: methods and applications*. 19. ed. [S.l.]: Social network analysis: methods and applications, 2009. Citado 5 vezes nas páginas 14, 21, 22, 24 e 25.

WASSERMAN, S.; GALASKIEWICZ, J. *Advances in social network analysis research in the social and behavioral sciences*. [S.l.]: SAGE, 1994. 299 p. Citado 3 vezes nas páginas 14, 21 e 22.

WHITE, H. D.; MCCAIN, K. W. Visualizing a discipline: An author co-citation analysis of information science 1972-1995. *Journal of the American Society for Information Science*, v. 49, n. 4, p. 327–355, 1988. Citado na página 36.

WU, T.; CHEN, Y.; HAN, J. Association mining in large databases: A re-examination of its measures. In: *Knowledge Discovery in Databases: PKDD*. [S.l.: s.n.], 2007. p. 621–628. Citado na página 52.

WU, T.; CHEN, Y.; HAN, J. Re-examination of interestingness measures in pattern mining: a unified framework. *Data Mining and Knowledge Discovery*, Kluwer Academic Publishers, Hingham, MA, USA, v. 21, n. 3, p. 371–397, 2010. Citado na página 52.

YANG, X.; STECK, H.; LIU, Y. Circle-based recommendation in online social networks. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2012. (KDD '12), p. 1267–1275. ISBN 978-1-4503-1462-6. Disponível em: <http://doi.acm.org/10.1145/2339530.2339728>. Citado na página 47.

YEUNG, K. F.; YANG, Y. A proactive personalized mobile news recommendation system. In: *Developments in E-systems Engineering (DESE), 2010*. [S.l.: s.n.], 2010. p. 207 –212. Citado na página 46.

ZHONG, E. et al. Modeling the dynamics of composite social networks. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2013. (KDD '13), p. 937–945. ISBN 978-1-4503-2174-7. Disponível em: <http://doi.acm.org/10.1145/2487575.2487652>. Citado na página 48.