UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

JAMISON JOSE DA SILVA LIMA

**Uma abordagem composta para a identificação de especialistas nas redes
sociais acadêmicas**

São Paulo

2017

JAMISON JOSE DA SILVA LIMA

**Uma abordagem composta para a identificação de especialistas nas redes
sociais acadêmicas**

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 12 de setembro de 2017. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Luciano Antonio Digiampietri

São Paulo

2017

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)

Lima, Jamison José da Silva

Uma abordagem composta para a identificação de especialistas nas redes sociais acadêmicas / Jamison José da Silva Lima ; orientador Luciano Antonio Digiampietri. – São Paulo, 2017.

82 f

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo.

Versão corrigida

1. Redes sociais. 2. Pesquisadores – Análise.
3. Currículo Lattes. I. Digiampietri, Luciano Antonio. II. Título.

CDD 22.ed.– 303.4833

Dissertação de autoria de Jamison Jose da Silva Lima, sob o título “**Uma abordagem composta para a identificação de especialistas nas redes sociais acadêmicas**”, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 12 de setembro de 2017 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Luciano Antonio Digiampietri

Universidade de São Paulo

Presidente

Prof. Dr. Leandro Augusto da Silva

Universidade Presbiteriana Mackenzie

Profa. Dra. Denise Hideko Goya

Universidade Federal do ABC

Prof. Dr. Edmir Parada Vasques Prado

Universidade de São Paulo

Aos meus pais.

Agradecimentos

Agradeço a todos que direta ou indiretamente me ajudaram a concluir esse projeto.

A minha família por todos esses anos que acreditaram no meu potencial e que apesar de todas as adversidades conseguiram me ajudar a me tornar a pessoa que sou.

A meus amigos pelo apoio emocional e por todos os momentos que me ajudaram a acreditar no meu potencial.

A todos meus professores pelo apoio acadêmico, em especial ao meu orientador Prof. Dr. Luciano Antonio Digiampietri, que me apresentou à ciência há cinco anos atrás.

A CAPES e a Universidade pelo apoio financeiro e estrutura essencial para minha formação acadêmica.

*“One machine can do the work of fifty ordinary men. No machine can do the work of one
extraordinary man”
(Elbert Hubbard)*

Resumo

LIMA, Jamison Jose da Silva. **Uma abordagem composta para a identificação de especialistas nas redes sociais acadêmicas**. 2017. 82 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2017. Versão corrigida.

A identificação de especialistas consiste na análise automática de informações sobre um conjunto de indivíduos para a localização daqueles com maior conhecimento em determinado tópico. Seus conhecimentos podem ser aplicados para a melhoria da produtividade em projetos de trabalho ou pesquisa e na localização de pesquisadores capacitados para avaliarem trabalhos de pesquisa ou para compor bancas de mestrado e doutorado. Para isso é interessante uma boa fonte de informações sobre os candidatos, de forma a otimizar o processo de identificação. No Brasil existe uma base de currículos que centraliza as informações sobre os pesquisadores brasileiros, a Plataforma Lattes. Os dados dessa plataforma são utilizados nesse trabalho para a realização da identificação de especialistas de acordo com suas áreas de atuação. São coletados dados da Plataforma e realizadas três etapas para compor uma lista ordenada dos principais especialistas de determinada área identificados pelo algoritmo proposto. A primeira etapa é a utilização da métrica TF-IDF, de modo que seus resultados são combinados com os estratos Qualis das publicações de cada autor, para finalmente ser aplicado um algoritmo de propagação para a geração da lista de especialistas. Após uma extensiva otimização dos parâmetros envolvidos no algoritmo, os resultados foram comparados com outra técnica existente que foi utilizada como base para esse projeto. Os resultados foram validados a partir de três conjuntos de pesquisadores coletados para esse projeto. O primeiro deles envolveu os pesquisadores contemplados com a Bolsa Produtividade em Pesquisa CNPq. Os resultados de dois questionários envolvendo pesquisadores das áreas de Inteligência Artificial e Nanotecnologia também foram utilizados. A partir dos resultados obtidos, foi possível notar que a técnica proposta aprimora significativamente os resultados do conjunto de Bolsistas Produtividade. Por outro lado, os resultados obtidos para os conjuntos de pesquisadores de Inteligência Artificial e Nanotecnologia foram ligeiramente melhores para a abordagem base. Os dados, algoritmos e resultados obtidos nesse trabalho poderão ser utilizados futuramente para aprimorar a identificação de especialistas.

Palavras-chaves: Identificação de especialistas. Redes sociais. Currículos Lattes.

Abstract

LIMA, Jamison Jose da Silva. **A composite approach for expert finding in academic social networks**. 2017. 82 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2017. Revised version.

Expert finding consists in the automatic analysis of informations about a set of people with the objective of finding those with greater knowledge on a particular topic. His knowledge can be applied to improve the productivity in research and work projects and to locate researchers who are capable of evaluating research papers or for masters and doctoral degrees. For this objective it's interesting a good source of information about the candidates, in order to optimize the process of identification. In Brazil there's a curriculum base that centralizes the information about Brazilian researchers, the Lattes Plataform. It's data is used in this work to perform the expert finding according to the candidate's areas of activity. Data are collected from the Platform and three steps performed to generate a ordered list of the main specialists of a given area identified by the proposed algorithm. The first step is to use the TF-IDF metric, so that its results are combined with the Qualis strata from the publications of each author, to finally be applied a propagation algorithm for the generation of the specialists list. After an extensive optimization of the parameters involved in the algorithm, the results were compared with another technique that was used as the basis for this project. The results were validated from three sets of researchers collected for this project. The first one involved the researchers contemplated with the CNPq Productivity Grant. The results of two questionnaires involving researchers from the areas of Artificial Intelligence and Nanotechnology were also used. From the results obtained, it was noticeable that the proposed technique significantly improves the results of the Productivity Grant set. On the other hand, the results obtained for the Artificial Intelligence and Nanotechnology research groups were slightly better for the base approach. The data, algorithms and results obtained in this work may be used in the future to improve the expert finding.

Keywords: Expert finding. Social networks. Lattes Curricula.

Lista de figuras

Figura 1 – Representação gráfica de um grafo não-direcionado	20
Figura 2 – Cálculo simplificado do PageRank	23
Figura 3 – Diagrama da extração de dados dos currículos Lattes	24
Figura 4 – Distribuição dos artigos de acordo com a estratégia	39
Figura 5 – Diagrama de dados do algoritmo desenvolvido	41
Figura 6 – Resultados das métricas AP@10 e AP@25 dada a variação do tamanho da lista de palavras	54
Figura 7 – Variação do R-Precision de acordo com o tamanho da lista de palavras	55
Figura 8 – Resultados para as métricas R-Precision e AP@25 dada a variação dos parâmetros do Score Beta	58
Figura 9 – Resultados a partir da variação dos parâmetros d e ϵ para a Computação	63
Figura 10 – Resultados a partir da variação dos parâmetros d e ϵ para a Química	64
Figura 11 – Resultados a partir da variação dos parâmetros d e ϵ para a Física	65

Lista de algoritmos

Algoritmo 1 – Estrutura básica do algoritmo de identificação de especialistas proposto . .	42
Algoritmo 2 – Inicialização do grafo	43
Algoritmo 3 – Inicialização dos dados das publicações para cada autor	44
Algoritmo 4 – Inicialização da lista de áreas de atuação dos pesquisadores	44
Algoritmo 5 – Criação da lista de palavras para uma área	45
Algoritmo 6 – Cálculo do Score Alfa	45
Algoritmo 7 – Cálculo do Score Propagated	48

Lista de tabelas

Tabela 1 – Classes de estratégias e métodos para validação de dados usados em cada artigo	40
Tabela 2 – Qualis da revista Computers in Biology and Medicine	46
Tabela 3 – Pesos atribuídos aos estratos Qualis de cada publicação	47
Tabela 4 – Valores padrão dos parâmetros utilizados nos testes preliminares	51
Tabela 5 – Resultados da identificação de especialistas em Ciência da Computação para o Score Alfa	52
Tabela 6 – Precisão média em relação ao número máximo de áreas permitidas para inclusão de termos na lista de palavras	53
Tabela 7 – Precisão e sensibilidade para os melhores tamanhos da lista de palavras	56
Tabela 8 – Valores padrão dos parâmetros utilizados nos testes do Score Beta . . .	57
Tabela 9 – Resultados parciais da identificação de especialistas em Ciência da Computação envolvendo o Score Beta	57
Tabela 10 – Valores padrão dos parâmetros utilizados nos testes do Score Propagated	59
Tabela 11 – Resultados parciais da identificação de especialistas em Ciência da Computação envolvendo o Score Propagated	60
Tabela 12 – Variação da precisão dos resultados de acordo com o tipo de média utilizada	61
Tabela 13 – Variação na precisão dos resultados a partir do parâmetro <i>subrede</i> . . .	62
Tabela 14 – Valores dos parâmetros utilizados após o treinamento	66
Tabela 15 – Lista de variações dos parâmetros para a realização dos próximos testes	67
Tabela 16 – Average Precision para cada conjunto de parâmetros com a normalização por área ativada	68
Tabela 17 – Average Precision para cada conjunto de parâmetros com a normalização por área desativada	68
Tabela 18 – Lista dos parâmetros selecionados após a otimização	69
Tabela 19 – Resultados para o conjunto de Bolsistas Produtividade	70
Tabela 20 – Resultados para o conjunto de Inteligência Artificial	71
Tabela 21 – Taxa de acertos da posição relativa para o conjunto de Nanotecnologia	72

Tabela 22 – Resultados para o conjunto de Nanotecnologia com a lista de palavras gerada <i>a priori</i>	73
Tabela 23 – Resultados para o conjunto de Nanotecnologia com a lista de palavras obtida automaticamente	73
Tabela 24 – Average Precision para cada composição do conjunto de Scores	75

Sumário

1	Introdução	15
1.1	<i>Objetivos</i>	16
1.2	<i>Metodologia</i>	16
1.3	<i>Organização do trabalho</i>	19
2	Conceitos fundamentais	20
2.1	<i>Grafos</i>	20
2.2	<i>Redes sociais</i>	21
2.3	<i>Algoritmos de propagação</i>	21
2.3.1	PageRank	22
2.3.2	HITS	22
2.4	<i>Mineração de textos</i>	24
2.5	<i>Técnicas de extração de informações</i>	25
2.6	<i>Identificação de especialistas</i>	26
3	Revisão bibliográfica	28
3.1	<i>Metodologia da revisão</i>	28
3.2	<i>Estratégias de identificação de especialistas</i>	29
3.2.1	Estratégias baseadas em perfil	30
3.2.2	Estratégias baseadas na estrutura da rede	30
3.2.3	Estratégias baseadas em conhecimento	31
3.2.4	Estratégias mistas	32
3.3	<i>Técnicas de avaliação</i>	36
3.4	<i>Considerações sobre a revisão</i>	38
4	Solução proposta	41
4.1	<i>Informações e estruturas utilizadas</i>	41
4.2	<i>Algoritmos implementados</i>	42
4.2.1	Inicialização do grafo	43
4.2.2	Inicialização das informações dos pesquisadores	43
4.2.3	Cálculo do Score Alfa	44

4.2.4	Cálculo do Score Beta	46
4.2.5	Cálculo do Score Propagated	47
4.3	<i>Considerações</i>	48
5	Resultados	50
5.1	<i>Dados coletados</i>	50
5.2	<i>Otimização dos parâmetros e resultados preliminares</i>	50
5.2.1	Score Alfa	51
5.2.2	Score Beta	56
5.2.3	Score Propagated	59
5.2.4	Comparação dos conjuntos de parâmetros	66
5.3	<i>Resultados finais</i>	69
5.3.1	Resultados do conjunto de Bolsistas Produtividade	70
5.3.2	Resultados do conjunto de Inteligência Artificial	71
5.3.3	Resultados do conjunto de Nanotecnologia	71
5.4	<i>Considerações</i>	74
6	Conclusões	76
6.1	<i>Contribuições</i>	76
6.2	<i>Trabalhos futuros</i>	77
	Referências¹	78

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

1 Introdução

A rede mundial de computadores possui o maior repositório de informação e conhecimento disponível publicamente (GOH; FUNG, 2005), possuindo bilhões de páginas indexadas na Internet (BOSCH; BOGERS; KUNDER, 2016). Muitas dessas informações podem ser analisadas e utilizadas para a obtenção de conhecimento, como trabalhos acadêmicos, notícias, perfis pessoais e livros. Um desses conjuntos é a Plataforma Lattes¹, que pode ser considerada como a principal base de dados envolvendo a pesquisa realizada no Brasil.

A Plataforma Lattes possui mais de 4,5 milhões de pesquisadores cadastrados, tendo entre eles mais de um milhão de currículos de mestres, doutores e especialistas (DIGIAMPIETRI et al., 2012; MENA-CHALCO et al., 2014; DIGIAMPIETRI et al., 2015; CNPq, 2016). Essa base foi criada para centralizar as informações dos currículos, grupos de pesquisa e instituições brasileiras, e seus dados podem ser utilizados em análises sobre seus pesquisadores e publicações.

Entre as aplicações possíveis para os dados da Plataforma Lattes estão a geração e análise de redes sociais a partir das coautorias dos autores (DIGIAMPIETRI et al., 2012; DIGIAMPIETRI et al., 2015), predição de novas relações de coautoria (DIGIAMPIETRI et al., 2015), análise de produtividade de grupos e pesquisa (LAENDER et al., 2008) e a identificação de especialistas a partir das relações de coautoria dos pesquisadores (CHAGAS; PEREZ-ALCAZAR; DIGIAMPIETRI, 2015), que foi o foco do presente trabalho.

A identificação de especialistas tem como objetivo localizar entre um conjunto de candidatos aqueles com maior conhecimento em determinado tópico a partir de suas informações pessoais, relacionamentos interpessoais e textos publicados (JOUNG et al., 2013; MITRE, 2006; ZHANG; TANG; LI, 2007).

A principal motivação para a identificação de especialistas é que sua presença pode aumentar a eficiência de projetos de trabalho ou pesquisa devido seus conhecimentos e pela capacidade de facilitar e aprimorar o trabalho de seus colegas (RIZOVA, 2008). Além disso, especialistas também são requeridos para a participação em bancas de defesa de mestrado ou doutorado, concursos públicos, bem como para a avaliação de artigos e projetos de pesquisa. A principal motivação para o desenvolvimento de sistemas automáticos para a identificação de especialistas é a dificuldade de encontrá-los manualmente (MITRE, 2006),

¹ lattes.cnpq.br

em particular se a busca for realizada por alguém sem conhecimento prévio sobre o tópico pesquisado.

Existem diversas aplicações possíveis para a identificação de especialistas, que vão desde a recomendação de pessoas adequadas para responder a questões em redes sociais online (SOUZA et al., 2013; BOK et al., 2015; ZHAO et al., 2015), a medição da qualidade de uma informação (KARDAN; MOHTAJ, 2013; BOK et al., 2015), encontrar colegas de trabalho com determinado conhecimento para resolverem problemas específicos (SILVA et al., 2012; CHEN et al., 2011) e a decisão de quais colaboradores serão convidados para opinar em determinado assunto (BARBOSA; SOUZA; OLIVEIRA, 2014), sendo esses apenas alguns exemplos das aplicações possíveis dessa área.

Devido a baixa quantidade de especialistas numa população, a identificação desses indivíduos torna-se um problema complexo, sendo assim um desafio apontá-los.

1.1 *Objetivos*

O objetivo geral desse projeto foi estender trabalhos prévios para a identificação de especialistas, aprimorando e generalizando sua execução a partir dos dados disponíveis na Plataforma Lattes, por meio de técnicas de extração de dados e análise de redes sociais, de forma a produzir uma solução mais robusta e precisa.

Os objetivos específicos desse trabalho foram os seguintes:

- Verificação do estado da arte para a identificação de especialistas nas redes sociais.
- Aprimoramento da técnica de identificação de especialistas proposta por Chagas, Perez-Alcazar e Digiampietri (2015).
- Otimização e análise dos parâmetros propostos para o algoritmo.

1.2 *Metodologia*

A metodologia dessa pesquisa iniciou-se pela realização de uma revisão sistemática envolvendo as bases digitais de artigos ACM² e IEEE Xplore³ para que fosse verificado o estado da arte para a identificação de especialistas nas redes sociais, suas principais aplicações e métodos de avaliação. Após o final da revisão foi detectado que seria possível

² dl.acm.org

³ ieeexplore.ieee.org

aprimorar a identificação de especialistas utilizando dados minerados dos currículos de pesquisadores da Plataforma Lattes, sendo assim proposta uma variação na técnica elaborada por Chagas, Perez-Alcazar e Digiampietri (2015).

Em seu algoritmo são realizadas duas etapas de inicialização antes da execução de um algoritmo de propagação, sendo elas o cálculo de uma métrica para análise da relevância dos títulos das publicações de cada periódico científico (denominado *Score Alfa*) e de uma métrica baseada nas avaliações Qualis de suas publicações (chamada de *Score Beta*). As avaliações Qualis são utilizadas para classificar a produção científica dos programas de pós-graduação por área de atuação, de modo que uma mesma publicação pode receber avaliações diferentes de acordo com a área, pois cada uma das áreas realiza avaliações distintas (Capes, 2016). Os resultados desses Scores são combinados e repassados como entrada para o cálculo de um algoritmo de propagação em grafos (ZHANG; TANG; LI, 2007), sendo o resultado dessa etapa chamado de *Score Propagated*.

O aprimoramento desse algoritmo foi realizado a partir de mudanças na lógica do cálculo dos Scores e na realização de diversos experimentos para identificar os principais fatores na otimização da identificação de especialistas. No algoritmo proposto nesse trabalho, os Scores Alfa e Beta são calculados levando em conta a lista de publicações de cada autor, a partir da hipótese que essa abordagem pode aprimorar os resultados obtidos pelo algoritmo já existente.

Para o estudo de caso realizado nesse trabalho, foram extraídos dados de três subconjuntos de pesquisadores da Plataforma Lattes. O primeiro deles consiste nos 13.358 doutores que recebiam Bolsa de Produtividade em Pesquisa do CNPq em 2016. O motivo para essa seleção foi de incluir somente pesquisadores com alto grau de contribuição científica, que é uma das exigências para o recebimento dessa modalidade de bolsa (CNPq, 2015). Essas bolsas são distribuídas e divididas por 76 conselhos de área, que atribuem critérios para divisão dessas bolsas de acordo com os principais fatores de relevância para a área. Existem dois níveis de bolsistas produtividade, sendo que o nível 1 contém os pesquisadores melhor avaliados em relação ao nível 2. Desse modo, os 5.926 pesquisadores que recebem a bolsa produtividade nível 1 foram considerados os especialistas do conjunto de dados nos testes realizados, que serão descritos em detalhes no capítulo 5.

Além dos dados envolvendo os Bolsistas Produtividade, foram coletados dados de outros dois conjuntos de pesquisadores cadastrados na Plataforma Lattes, sendo um deles formado por 3.343 pesquisadores que declararam ser da área de Inteligência Artificial (IA),

por meio de uma lista de palavras-chave relacionadas à área. Essa seleção foi realizada a partir da busca no banco de dados do projeto por pesquisadores que declararam no campo “Áreas de Atuação” de seus currículos Lattes atividades referentes à área de Inteligência Artificial. A lista de termos utilizados na busca incluiu os seguintes tópicos (permitindo-se pequenas variações na grafia): algoritmos evolutivos, aprendizado de máquina, *big data*, computação evolutiva, computação inteligente, engenharia de conhecimento, inteligência artificial, inteligência computacional, lógica nebulosa, mineração de dados, processamento de linguagem natural, reconhecimento de padrões, redes Bayesianas, redes complexas, redes neurais e sistemas inteligentes.

Para a validação dos resultados desse conjunto foi realizado um questionário no qual foi pedido a pesquisadores da área que listassem entre 1 e 5 colegas da área que considerassem especialistas, de acordo com seu conhecimento sobre a área. Esse questionário foi respondido por 11 pesquisadores diferentes, e a partir de suas respostas foram obtidos 25 nomes de pesquisadores, sendo que 24 deles possuem currículos cadastrados na Plataforma Lattes, sendo esses 24 considerados os especialistas para os testes realizados nesse conjunto.

Por último foram reaproveitados os dados coletados por Chagas, Perez-Alcazar e Digiampietri (2015) envolvendo 1.067 pesquisadores em Nanotecnologia, dos quais 155 foram removidos dessa lista para os testes realizados nesse trabalho por não possuírem coautorias ou publicações no banco de dados utilizado nesse projeto. Além disso foi obtida a ontologia utilizada em seu trabalho (ALUÍSIO et al., 2005) e os resultados do questionário realizado, consistindo em 66 comparações entre pares de pesquisadores da área envolvendo assuntos relacionados à nanotecnologia, como nanopartículas e ácidos. Em seu experimento, foram realizados questionários perguntando a pesquisadores da área qual entre dois de seus colegas possuía maior conhecimento em determinado tópico relacionado à Nanotecnologia. A partir desses resultados foram realizados testes limitando a ontologia para cada tópico e comparando a classificação dos pesquisadores envolvidos no algoritmo com os resultados obtidos pelos questionários respondidos. Além disso foi gerada uma lista de especialistas envolvendo todos os 27 pesquisadores indicados por seus conhecimentos nos questionários realizados.

No total foram obtidos dados de 17.387 pesquisadores (pois alguns deles estão em mais de um conjunto) e cerca de 1.2 milhão de títulos de publicações a partir de uma versão atualizada do banco de dados criado por Digiampietri et al. (2014), incluindo dados

de até julho de 2015 envolvendo as publicações, coautorias, áreas de atuação e o Qualis para cada uma das áreas.

A avaliação do desempenho foi realizada utilizando diferentes medidas de acurácia e revocação, que serão detalhadas ao longo do texto.

1.3 Organização do trabalho

O restante desta dissertação está organizado da seguinte forma. O capítulo 2 descreve os conceitos básicos para o entendimento deste projeto. O capítulo 3 relata os trabalhos relacionados ao tema, com destaque para as abordagens utilizadas para a identificação de especialistas e seus métodos de avaliação. O capítulo 4 apresenta a solução proposta. O capítulo 5 aponta como foi realizada a otimização do algoritmo e discute os resultados obtidos. Por fim, o capítulo 6 apresenta as conclusões e contribuições do projeto além de sugestões para trabalhos futuros.

2 Conceitos fundamentais

Neste capítulo são descritos os conceitos básicos utilizados nesse projeto, sendo eles grafos, redes sociais, algoritmos de propagação, mineração de textos, Plataforma Lattes, técnicas de recuperação de informações e identificação de especialistas.

2.1 Grafos

Um grafo não-direcionado é um conjunto $G(V, E)$, onde V é um conjunto não vazio de vértices e E é um conjunto de arestas, sendo que cada aresta consiste num par não ordenado de vértices, indicando que há algum tipo de relação entre eles. Por exemplo, em um mapa cada vértice pode consistir em uma cidade e cada aresta numa estrada conectando-as. Em uma rede social cada vértice pode consistir em uma pessoa e cada aresta num relacionamento de amizade entre elas. A figura 1 é um exemplo de grafo não-direcionado contendo 6 vértices e 7 arestas.

Figura 1 – Representação gráfica de um grafo não-direcionado

Fonte: Jamison Jose da Silva Lima, 2017

Um grafo pode também possuir relações direcionadas, de modo que cada aresta consiste num par ordenado de vértices, indicando uma relação de um vértice para o outro. Esse tipo de grafo é denominado digrafo ou grafo direcionado, sendo útil para representar, por exemplo, redes de *hyperlinks* entre páginas da rede mundial de computadores ou de citações entre artigos. Outra variação comum são os grafos com atributos, nos quais tanto os vértices quanto as arestas podem conter atributos que ajudem a representá-los, como a distância entre duas cidades ou o número de citações recebidas por cada artigo. Para

este trabalho foram criadas redes sociais representadas por grafos não-direcionados com atributos, como será apresentado a seguir.

2.2 *Redes sociais*

As redes sociais são estudadas há muito tempo no campo das ciências sociais e comportamentais devido aos padrões que podem ser encontrados nessa estrutura (WASSERMAN; FAUST, 1994). Para o contexto deste trabalho, redes sociais podem ser definidas como uma rede contendo atores e seus relacionamentos (JOUNG et al., 2013). Uma classe de redes sociais muito popular atualmente são as redes sociais virtuais, consistindo estruturalmente por relações de amizade entre seus usuários. Existem também redes nas quais seus atores são documentos relacionados por meio de citações (PAGE et al., 1999; KLEINBERG, 1999).

Para este trabalho foram criadas redes sociais de coautorias (NEWMAN, 2001) envolvendo pesquisadores brasileiros cadastrados na Plataforma Lattes, em que cada relacionamento entre pares de pesquisadores indica que existe pelo menos um trabalho publicado em conjunto por eles, e quanto maior o número de coautorias, maior o peso da relação na rede.

Um dos principais motivos para a utilização de redes sociais na identificação de especialistas é o conceito de homofilia, que consiste na suposição de que um indivíduo possui maior probabilidade de se relacionar com pessoas semelhantes (ZHANG et al., 2012; LI et al., 2013). A partir desse conceito é possível inferir que há uma maior probabilidade de especialistas estarem mais próximos numa rede social.

Existem diversas técnicas para analisar as características e modelar as relações entre os atores da rede. Entre elas estão os algoritmos de propagação, que serão detalhados na próxima seção.

2.3 *Algoritmos de propagação*

A estrutura de redes sociais é comumente utilizada para a identificação de especialistas devido aos algoritmos de propagação, que levam em consideração a influência de cada ator calculada a partir da quantidade e qualidade de seus relacionamentos. Entre

os algoritmos de propagação existentes, dois dos primeiros e mais importantes foram propostos por Page et al. (1999) e Kleinberg (1999), denominados respectivamente como PageRank e HITS. É relevante notar que esses algoritmos foram inicialmente criados com o objetivo de aprimorar os resultados de busca na rede mundial de computadores, e que somente depois de alguns anos começaram a ser adaptados para a identificação de especialistas, devido a correlação entre nível de especialidade de um indivíduo e a qualidade de uma página. Além disso, as redes sociais possuem intrinsecamente o princípio de homofilia por representarem as relações entre os indivíduos.

2.3.1 PageRank

A estrutura original do PageRank consiste na rede de citações criada a partir dos *hyperlinks* de cada página da Internet, para então ser calculada a relevância de cada página a partir da quantidade e qualidade de suas citações, sendo a última medida pela importância da página que realiza a citação. Seguindo essa lógica, as páginas mais importantes são referenciadas por outras também relevantes.

O PageRank é calculado da seguinte forma: inicialmente, todos os nós da rede recebem o mesmo valor, sendo este atualizado para todos os nós da rede por meio da fórmula apresentada na equação 1.

$$PR(u) = c \sum_{\forall v \in Bu} \frac{PR(v)}{Nv} + cE(u), \quad (1)$$

onde Bu é a lista de vizinhos do nó u , Nv é o número de vizinhos do nó v , $E(u)$ é o valor inicial de u e c é um fator de propagação. Esse algoritmo é executado iterativamente até a convergência.

A figura 2 apresenta um exemplo simples para o algoritmo já executado até a convergência, de modo que a página B possui o menor valor por receber apenas uma citação da página A, sendo que A repassa parte de seu valor tanto para A quanto para C.

2.3.2 HITS

O algoritmo HITS tem como objetivo localizar as páginas mais relevantes em determinado tópico. Para isso, é selecionado um subconjunto de páginas da Internet (selecionadas por sua relevância), sendo que originalmente eram utilizados resultados de

Figura 2 – Cálculo simplificado do PageRank

Fonte: Page et al. (1999)

páginas de busca para a geração desse subconjunto. Esse conjunto extraído inicialmente é denominado *raiz* e ele é então expandido por meio dos *hiperlinks* existentes envolvendo a raiz. Essas páginas adicionadas formam o conjunto *base*, que é o utilizado no cálculo do HITS.

Após a criação do conjunto base são calculados os valores denominados como pivô e autoridade para cada página, sendo que o primeiro representa a quantidade de referências a outras páginas e o segundo o nível de autoridade calculado pela quantidade e qualidade das referências recebidas. A autoridade p_{auth} para a página p é calculada iterativamente utilizando da equação 2:

$$p_{auth} = \sum_{q:(q,p) \in E} q_{hub}, \quad (2)$$

onde $\{q : (q, p) \in E\}$ é o conjunto de páginas com *hiperlinks* para p , e q_{hub} é o valor pivô referente a página q .

O pivô é calculado de maneira análoga, como apresentado na equação 3:

$$p_{hub} = \sum_{q:(p,q) \in E} q_{auth}, \quad (3)$$

sendo $\{q : (p, q) \in E\}$ o conjunto de páginas citadas por p .

As Equações 3 e 2 são aplicadas a todo o subconjunto selecionado por um número determinado de iterações ou até sua convergência.

Figura 3 – Diagrama da extração de dados dos currículos Lattes

Fonte: Digiampietri et al. (2014)

2.4 Mineração de textos

O grupo de pesquisa no qual este projeto está contextualizado mantém uma base de dados de currículos da Plataforma Lattes (DIGIAMPIETRI et al., 2012; MENA-CHALCO et al., 2014; DIGIAMPIETRI et al., 2015). Atualmente, esta base conta com quase 5 milhões de currículos de pesquisadores, e a última atualização desse conjunto foi realizada em julho de 2015.

O processo atual para a geração da base de dados é o seguinte (DIGIAMPIETRI et al., 2014): para cada pesquisador é inicialmente localizado seu currículo e convertido para o formato XML. É então realizada a extração de dados a partir de cada arquivo XML, utilizando expressões regulares para obter informações pessoais, como nome e local de trabalho, e também são obtidos dados sobre suas publicações em revistas e conferências, incluindo nome da revista, coautores, ano e título, que são utilizados para a criação de um banco de dados. Então são obtidas métricas externas envolvendo as publicações, como os fatores de impacto JCR e SJR, o número de citações e os índices G e H são obtidos dos sites Microsoft Academic Search¹ e Google Scholar², e por último o Qualis por área para cada publicação. O Qualis é dividido nos conceitos A1 e A2, B1 a B5 e C, e como cada área realiza uma avaliação diferente para essa métrica, é possível comparar a importância das publicações em cada uma das 51 áreas definidas para essa métrica. A figura 3 apresenta o diagrama desse processo.

Nesse projeto foram utilizados dados envolvendo as publicações dos autores, suas áreas de atuação, coautorias e a métrica Qualis para identificação dos especialistas.

¹ academic.research.microsoft.com

² scholar.google.com

2.5 Técnicas de extração de informações

Diversas abordagens para a identificação de especialistas utilizam técnicas de extração de informações como forma de inferir o conhecimento dos candidatos a partir de suas publicações. A ideia geral desses sistemas é estimar a especialidade de cada candidato calculando a relação entre seus documentos textuais e o tópico que estão sendo buscados os especialistas.

A técnica *Term Frequency-Inverse Document Frequency* (TF-IDF) (PUTTAN; SENIVONGSE, 2013) é uma medida estatística que busca a relevância de cada documento a partir do raciocínio que os termos relevantes com alta frequência em determinado documento e que não aparecem em outros são considerados mais importantes para a caracterização do documento em relação aos demais, como mostrado na equação 4:

$$TF_{ij} * IDF_i = \frac{n_{ij}}{\sum k n_{kj}} * \log \frac{|D|}{|\{d : t_i \in d\}|}, \quad (4)$$

onde n_{ij} é a frequência do termo i no documento j , $|D|$ é o tamanho do conjunto de documentos e $|\{d : t_i \in d\}|$ é o número de documentos em que o termo t_i aparece.

Caso seja necessário calcular o TF-IDF para um conjunto de termos relevantes (como acontece nesse trabalho), pode ser calculada a soma do TF-IDF para cada um dos termos relevantes em relação ao documento.

Um dos principais problemas desse método é a dificuldade em se localizar especialistas caso os documentos não forem totalmente relacionados com o tópico, pois essa técnica analisa apenas a frequência dos termos relevantes em relação aos encontrados no conjunto de documentos. Apesar disso, TF-IDF é uma métrica de implementação consideravelmente simples e relevante para a recuperação de informação.

Existem outras técnicas que utilizam modelos estatísticos de tópicos para a identificação de especialistas, sendo essas técnicas que levam em consideração tanto a sintaxe quanto a semântica para a recuperação de informações (SRBA; GRZNAR; BIELIKOVA, 2015). Nessa classe de modelos existem dois frequentemente utilizados na identificação de especialistas, o *Explicit Semantic Analysis* (ESA) e o *Latent Dirichlet Allocation* (LDA). Li et al. (2013) utilizam a técnica ESA para a comparação entre documentos usando conceitos extraídos da Wikipédia por exemplo. Já o modelo LDA utiliza a distribuição multivariada Dirichlet para inferir a qualidade de um documento com relação a cada tópico (SRBA; GRZNAR; BIELIKOVA, 2015).

2.6 Identificação de especialistas

Um especialista pode ser considerado como um indivíduo com a capacidade de responder questões complexas com qualidade (MOVSHOVITZ-ATTIAS et al., 2013). A identificação de especialistas consiste em encontrar a partir de uma lista de candidatos aqueles que possuam maior conhecimento em determinado tópico. Existem diversas técnicas na literatura que podem realizar essa identificação, e neste trabalho as estratégias existentes serão classificadas a partir das categorias de informação que são utilizadas: os perfis dos candidatos, atributos de redes sociais e o conhecimento dos atores (JOUNG et al., 2013). Essa seção apresenta uma análise geral para a identificação de especialistas. Os detalhes sobre cada uma das estratégias estudadas serão detalhadas no capítulo 3. Já a solução proposta será descrita no capítulo 4.

Uma das primeiras e mais conhecidas abordagens para a identificação de especialistas nas redes sociais é a proposta por Campbell et al. (2003), denominada *email expertise extraction* (e^3), que é iniciada pela extração das informações sobre determinado tópico por meio de um servidor de emails extraídos por meio de palavras-chave e clusterização não-supervisionada, para então ser gerada uma rede social a partir dos remetentes e destinatários dos emails, e finalmente ser aplicado um algoritmo baseado no HITS (KLEINBERG, 1999) para classificar os candidatos. Campbell et al. (2003) também exploram brevemente as origens da identificação de especialistas, que antes de utilizarem a estrutura de redes sociais consistiam exclusivamente na extração de informações de perfil ou de seus documentos para inferir o conhecimento de cada candidato.

A técnica proposta nesse trabalho utiliza uma estrutura semelhante, porém aprimorada da técnica e^3 , realizando a recuperação de informações a partir do cálculo do TF-IDF e pelos estratos Qualis, para então ser aplicado uma adaptação do algoritmo de propagação proposto por Zhang, Tang e Li (2007), que será detalhado na Seção 3.2.4.

Existem outras estratégias que utilizam métodos estatísticos ou matemáticos para a identificação de especialistas, como o modelo proposto por (ZHANG et al., 2012) que cria um sistema complexo onde cada participante da rede possui informações locais envolvendo apenas seus vizinhos na rede, de modo que a similaridade entre autores é calculada por meio de uma regressão logística.

No capítulo 3 são descritas as técnicas para a identificação de especialistas encontradas na revisão bibliográfica realizada.

3 Revisão bibliográfica

Este capítulo possui como objetivo sintetizar a pesquisa bibliográfica sobre identificação de especialistas realizada por meio de uma revisão sistemática, sendo esse um método de pesquisa que realiza uma busca extensa e sistemática sobre o trabalho realizado em uma área.

3.1 Metodologia da revisão

Uma revisão sistemática possui três etapas, sendo elas: planejamento, condução e documentação. Na primeira fase é identificado o objetivo da revisão e é desenvolvido seu protocolo. Durante a segunda fase é formulada e executada a busca por publicações, os trabalhos relevantes são selecionados, avaliados e seus dados são extraídos e sintetizados. A terceira fase consiste na publicação dos resultados.

Objetivo: Estudar o estado da arte na identificação de especialistas nas redes sociais.

As questões da pesquisa foram as seguintes:

Questão primária: “Quais são as abordagens possíveis para a identificação de especialistas nas redes sociais?”

Questão secundária: “Quais são os métodos e técnicas de avaliação utilizados para comparar os algoritmos de identificação de especialistas?”

Os recursos de busca e seleção de estudos foram os seguintes:

Identificação de fontes de busca: Foram realizadas pesquisas nas bases de publicações IEEE Xplore e ACM.

Idioma dos trabalhos: Inglês.

Data dos trabalhos: Os artigos selecionados foram publicados entre 1º de janeiro de 2011 e 31 de dezembro de 2015.

Critérios de inclusão:

- Possuir algoritmos de identificação de especialistas;
- Aplicar a identificação de especialistas para a resolução de um problema.

Critérios de exclusão:

- Não aplicar a estrutura de rede social para a identificação de especialistas;
- Apontar especialistas a priori;
- Ponderar as relações exclusivamente por distância.

Strings de busca na base de publicações IEEE Xplore:

- (((("Index Terms":expert finding) OR "Index Terms":expert search) AND "Index Terms":social network*));
- (((("Index Terms":“expert finding” OR “Index Terms”:“expert search” OR “Index Terms”:“expertise finding” OR “Index Terms”:“expertise search”) AND (“Index Terms”:“social network” OR “Index Terms”:“social networks”)))).

Strings de busca na base de publicações ACM:

- (+expert* +find* +social +network*);
- + (“expert finding” “expert search” “expertise finding” “expertise search”) + (“social network” “social networks”).

A partir do planejamento descrito, foram encontrados 97 artigos nas duas bases de dados, que formaram nossa seleção preliminar, e então foi iniciada a condução. Após a leitura de seus resumos e aplicação dos critérios de inclusão e exclusão, 36 foram pré-selecionados para a extração, sendo que 29 deles foram escolhidos para comporem a síntese da pesquisa após a leitura dos artigos e uma segunda aplicação dos critérios de inclusão e exclusão. Além dos artigos extraídos pela revisão sistemática, outros 6 foram incluídos manualmente por serem relevantes para a pesquisa mesmo estando fora do escopo de seleção, por serem mais antigos ou por estarem escritos em português ao invés do inglês que foi utilizado nessa revisão.

3.2 Estratégias de identificação de especialistas

Como já apresentado na seção 2.6, as estratégias existentes para identificação de especialistas podem ser categorizadas a partir das classes de informações coletadas: os perfis dos candidatos, atributos da rede e o conhecimento dos atores (JOUNG et al., 2013). Nessa seção são apresentadas as técnicas propostas pelos artigos pesquisados, e os métodos de avaliação apresentados por eles.

3.2.1 Estratégias baseadas em perfil

A classe mais simples de estratégias para identificação de especialistas consiste na coleta das informações de perfil dos candidatos, incluindo idade, gênero, interesses, localização e ocupações (JOUNG et al., 2013). Devido ao escopo da revisão sistemática, que limitou as abordagens de identificação de especialistas para incluir somente técnicas envolvendo redes sociais, foi encontrada apenas uma técnica que utiliza somente as informações de perfil.

O objetivo da técnica proposta por Zhang et al. (2012) é localizar o especialista com maior similaridade ao perfil ideal. Para isso é gerado um sistema complexo onde cada participante da rede possui informações locais envolvendo apenas seus vizinhos na rede, de modo que a similaridade entre autores é calculada por meio de uma regressão logística.

3.2.2 Estratégias baseadas na estrutura da rede

As estratégias que empregam a estrutura da rede na identificação de especialistas levam em consideração a quantidade e/ou a qualidade das relações entre os atores da rede para a classificação dos candidatos. Diversas técnicas aproveitam essa estrutura para aplicar algoritmos de propagação, geralmente por meio de técnicas semelhantes ao PageRank (PAGE et al., 1999) e ao HITS (KLEINBERG, 1999), em que cada ator da rede é classificado a partir do número e da qualidade de suas relações. A lógica por trás dessa classe de estratégias é que as relações de um ator podem indicar seu conhecimento e influência numa rede social devido a homofilia.

Foram encontradas nessa revisão quatro abordagens diferentes que aplicam apenas a estrutura de rede para a identificação de especialistas.

O algoritmo ExpertiseRank (ZHANG; ACKERMAN; ADAMIC, 2007) apresenta uma variação do PageRank em conjunto com uma métrica obtida em uma rede social formada a partir de um fórum da Internet, que leva em consideração para cada candidato a quantidade de respostas e uma métrica de qualidade da resposta de acordo com a quantidade de respostas dadas à questão.

A técnica SNPageRank (KARDAN; OMIDVAR; FARAHMANDNIA, 2011) busca aprimorar o PageRank para aplicação em redes sociais, de modo que cada relação passa

a possuir um peso indicando sua importância. Essa técnica foi aprimorada incluindo a direção e o fluxo das relações (KARDAN; MOHTAJ, 2013), a partir da hipótese que os atores no final de uma cascata de informação possuem maior influência na rede.

A técnica APINet (PADHYE; MUKHERJEE; SINHA, 2014) usa uma versão modificada do PageRank para gerar recomendações de especialistas a partir dos dados minerados da plataforma de desenvolvimento colaborativo GitHub¹. Além disso, são recomendados projetos e bibliotecas de acordo com os interesses do usuário.

O método proposto por Shahriari, Parekodi e Klamma (2015) aplica os algoritmos HITS e PageRank em conjunto com detecção de comunidades sobrepostas (do inglês *overlapping community detection*) para localizar especialistas em duas categorias do fórum Stack Overflow² e em um fórum sobre a vida selvagem estoniana³. O objetivo da detecção de comunidades sobrepostas é separar os usuários em subgrupos de acordo com cada tópico, de modo que apenas usuários ativos participem e possuam maior peso na rede para esse determinado tópico.

3.2.3 Estratégias baseadas em conhecimento

As estratégias de conhecimento usam os conteúdos publicados pelos candidatos como trabalhos acadêmicos, relatórios e postagens (JOUNG et al., 2013). Para isso são aplicadas técnicas de recuperação de informação para analisar os documentos dos candidatos e inferir seus conhecimentos.

O trabalho publicado por Bok et al. (2015) propõe uma técnica para a indicação de especialistas por meio da extração de postagens recentes em uma rede social, sendo elas classificadas de acordo com o peso de suas palavras-chave para o tópico e por sua popularidade. Após a classificação dos documentos os candidatos são ranqueados, levando em consideração a pontuação de suas publicações, seu nível de confiabilidade a partir das avaliações de outros usuários da rede e a probabilidade da questão ser respondida por ele.

O método proposto por Movshovitz-Attias et al. (2013) busca inferir entre os novos usuários do fórum Stack Overflow quais serão especialistas futuramente por meio da composição de métricas envolvendo suas atividades no site.

¹ github.com

² stackoverflow.com

³ looduskalender.ee

3.2.4 Estratégias mistas

Diversas técnicas aproveitam-se de mais do que uma classe de estratégias para a identificação de especialistas. Nessa subseção são descritas as técnicas encontradas nessa revisão que utilizam essas combinações, e como poderá ser notado, a maior parte delas aplica a estrutura de rede para esse fim.

Estratégias de perfil e estrutura da rede

Nessa revisão foi identificada uma técnica envolvendo informações de perfil em conjunto com a estrutura da rede social para a identificação de especialistas. No modelo proposto por Li et al. (2012) são usados dados de uma rede social corporativa, onde as especialidades de um usuário são inseridas manualmente. A partir das recomendações de especialidade entre os usuários são gerados grafos, sendo aplicado então um algoritmo de propagação semelhante ao PageRank para inferir os principais especialistas de cada área.

Estratégias de perfil e conhecimento

A partir da pesquisa realizada, foi encontrada uma abordagem que combina os perfis dos pesquisadores e seus conhecimentos para a identificação de especialistas. Na abordagem proposta por Srba, Grznar e Bielikova (2015), informações de perfil e publicações dos candidatos são mineradas da rede social Twitter⁴, por suas questões publicadas no fórum *Stack Overflow* e também pelas suas postagens em seus *sites* pessoais. Essas informações são aplicadas ao modelo LDA para a identificação dos especialistas.

Estratégias de estrutura de rede e conhecimento

As estratégias de identificação de especialistas muitas vezes utilizam técnicas de mineração de texto em conjunto a métricas de redes sociais para a identificação de especialistas.

No trabalho proposto por Campbell et al. (2003), os emails dos candidatos referentes ao tópico buscado são coletados e analisados para a geração de um grafo direcionado e

⁴ www.twitter.com

ponderado de candidatos a especialistas, para então ser aplicada uma versão modificada do algoritmo de propagação HITS.

O algoritmo proposto por Souza et al. (2013) utiliza como base a lista de seguidores do Twitter para identificar entre eles o mais apropriado para responder a uma questão. Para isso são calculados os seguintes critérios: conhecimento sobre o tópico, proximidade entre o usuário que está realizando a busca com relação a seus seguidores e tempo de resposta.

Liu et al. (2013) apresenta uma técnica de identificação de especialistas numa rede social corporativa que consiste na seleção de candidatos a partir de uma busca em determinado tópico, sendo então realizada a construção de dois índices. O primeiro índice refere-se às atividades do candidato com relação ao tópico, sendo ele composto pela taxa de postagens, comentários, pessoas que seguem e são seguidas pelo candidato. O segundo índice representa a influência e correlação com o tópico, por meio dos compartilhamentos e respostas para suas publicações, além da similaridade entre candidato e o tópico calculado pelo TF-IDF.

A técnica proposta por Chen et al. (2011) é iniciada pela identificação dos principais documentos referentes a um tópico por meio de técnicas de extração de informações, para então ser gerada uma rede envolvendo os autores citados nesses documentos. Após a criação da rede um algoritmo de propagação semelhante ao PageRank é utilizado, tendo como valores iniciais os resultados de um modelo de linguagem que leva em consideração a quantidade de vezes que cada autor é citado no conjunto de documentos.

A partir das hipóteses de que quanto mais uma pessoa participa nas discussões de um tópico, maior é seu interesse nele e de que um documento torna-se mais importante de acordo com a quantidade de encaminhamentos realizados por pessoas relevantes, Li et al. (2012) propõem a seguinte técnica para a identificação de especialistas numa rede de *blogs*: inicialmente é utilizando o modelo LDA para identificar o tópico de cada documento, então a especialidade dos candidatos é calculada por um algoritmo de propagação que envolve a qualidade de seus documentos, que é calculada pelas interações de seus leitores no documento e pela influência deles na rede. Além disso são contabilizadas as atividades do autor no tópico, de acordo com a quantidade de documentos postados e por seu interesse na leitura de artigos relacionados.

O método criado por Puttan e Senivongse (2013) analisa em um grupo do Facebook⁵ a quantidade de curtidas, o número de palavras, a complexidade (calculada pela quantidade de termos distintos) e a informatividade do comentário (calculada pelo TF-IDF). Esses valores formam o valor inicial que é aplicado em um algoritmo de propagação semelhante ao ExpertiseRank (ZHANG; ACKERMAN; ADAMIC, 2007).

A abordagem proposta por Yeniterzi e Callan (2014) é baseada no algoritmo HITS, com a principal diferença de que a rede criada contém somente postagens relevantes ao tópico, a partir da hipótese de que os candidatos podem ter publicações em diversos tópicos.

A técnica proposta por Zhao et al. (2015) utiliza técnicas de estimação de valores desconhecidos para modelar a probabilidade de cada candidato responder corretamente novas questões com base em suas respostas anteriores. Esses resultados são regularizados por meio de um grafo gerado pelas relações sociais dos usuários no Twitter, partindo da suposição de que se dois usuários são próximos na rede, eles possuem maior probabilidade de conseguirem responder as mesmas questões.

Deng et al. (2012) apresentam uma estrutura de rede social contendo atores, documentos e locais de publicação, levando em conta três restrições de regularização para o cálculo da identificação de especialistas: consistência de documento, onde documentos parecidos possuem relevância parecida; consistência de coautoria, se duas pessoas são coautoras em diversos trabalhos, possuirão conhecimento parecido; e a consistência entre autores e documentos, de modo que especialistas em determinado tópico tenham publicado artigos sobre ele e documentos relevantes sobre o tópico tenham sido publicados por especialistas da área.

Smirnova (2011) apresenta um modelo estatístico para a identificação de especialistas em determinado tópico, no qual inicialmente são criados perfis para os candidatos, onde estão inclusas as relações entre os atores, suas publicações e o tópico. Para a geração de uma rede social é utilizada a similaridade entre os perfis criados. A identificação dos especialistas é feita pelo cálculo da probabilidade conjunta dos termos do tópico serem conhecidos pelo autor.

Welser et al. (2011) apresentam uma técnica para identificar classes de funções dos editores da Wikipédia a partir da quantidade de participações em cada seção e conversas

⁵ www.facebook.com

com outros usuários, sendo que os especialistas foram detectados principalmente pela maior participação em páginas de discussão.

A proposta de Bozzon et al. (2013) cria um modelo envolvendo os usuários de redes sociais, detectando especialistas entre seus conhecidos próximos por meio dos termos relevantes em suas postagens e seus relacionamentos. Os candidatos são classificados por meio de uma variação da técnica TF-IDF para seus conteúdos na rede de acordo com a relevância da informação para o tópico buscado.

Estratégias de perfil, estrutura de rede e conhecimento

Algumas abordagens utilizam dados de perfil e publicações dos autores em conjunto com métricas de rede para a identificação de especialistas, como na abordagem proposta por Zhang, Tang e Li (2007), onde as informações sobre pesquisadores são coletadas para a geração de um perfil, incluindo informações pessoais, coautorias, número de artigos publicados e citações. A partir da relevância do perfil em relação ao tópico cada autor recebe um valor inicial para então ser aplicado o seguinte algoritmo de propagação:

$$s(v_i)^{n+1} = s(v_i)^n \sum_{v_j \in V_i} \sum_{e \in R_{ji}} w((v_j, v_i), e) s(v_j)^n, \quad (5)$$

onde $v_j \in V_i$ é a lista de vizinhos v_j de v_i , $e \in R_{ji}$ é a lista de tipos de relacionamento R_{ji} entre v_j e v_i , e $w((v_j, v_i), e)$ é o peso do tipo de relacionamento e entre os nós v_i e v_j .

O método proposto no trabalho de Li et al. (2013) envolve as publicações e interesses declarados juntamente com a estrutura da rede, sendo que seu conhecimento é mensurado pela técnica *Explicit Semantic Analysis* e utilizado em um algoritmo de propagação similar ao PageRank.

O algoritmo criado por Liou e Li (2015) tem como objetivo identificar especialistas em locais de entretenimento, compras e comida a partir da quantidade de postagens e visitas em determinada categoria, influência medida por sua centralidade e popularidade em redes sociais e proximidade do local onde está sendo realizada a busca.

A abordagem proposta por Chagas, Perez-Alcazar e Digiampietri (2015) é a base para o algoritmo proposto nesse trabalho, consistindo na obtenção de dados de pesquisadores que possuem currículos cadastrados na Plataforma Lattes e dados sobre o Qualis de suas publicações para a identificação de especialistas na área de Nanotecnologia. Pri-

meiramente é calculada a qualidade das publicações por meio de dados do Qualis e da aplicação do TF-IDF envolvendo os títulos das publicações para determinado assunto, levando em consideração uma ontologia referente à área. Após a geração dos índices para cada publicação, esses valores são repassados aos autores, sendo então usado o algoritmo de propagação proposto por Zhang, Tang e Li (2007) e descrito na equação 5 a partir da rede de coautorias obtida pela extração dos currículos Lattes.

3.3 Técnicas de avaliação

A avaliação das técnicas de identificação de especialistas geralmente é realizada a partir de uma lista classificada pela ordem ou pelo nível os candidatos existentes. Essas listas podem ser geradas por meio da classificação manual ou pela extração de informações dos candidatos.

Para a classificação manual geralmente são consultadas pessoas com conhecimento no assunto o qual os especialistas estejam sendo buscados, e a partir de seus conhecimentos os candidatos são classificados. As listas geradas por meio de técnicas de recuperação de informação utilizam os dados dos candidatos para a classificação dos especialistas.

A classificação dos resultados é realizada geralmente pela divisão binária entre especialistas e não-especialistas ou pelas classificações individuais dos candidatos. Também pode ser realizada a classificação por níveis de conhecimento, de modo que candidatos com o mesmo nível de conhecimento sejam agrupados, não importando a ordem interna do grupo.

As principais métricas de avaliação usadas para a identificação de especialistas são as seguintes (ZHAO et al., 2015; DENG et al., 2012; SU; YUAN; ZHU, 2015):

A métrica *Precision@n* ($P@n$) compara a precisão da lista obtida em relação a classificação ideal, levando em consideração os primeiros n classificados. A precisão é calculada como $P@n = \frac{\sum_1^n ex}{n}$, onde $\sum_1^n ex$ representa o número de especialistas identificados corretamente entre os n primeiros. Essa é uma das técnicas mais utilizadas pois a classificação dos especialistas é especialmente importante entre os primeiros resultados.

O complemento da precisão é a sensibilidade, denominada como *Recall@n* ($R@n$), consistindo na proporção do total de especialistas encontrados entre os n primeiros. A fórmula dessa métrica é $R@n = \frac{\sum_1^n ex}{total_{ex}}$, onde $total_{ex}$ é o número total de especialistas da

amostra. Como em diversos casos o tamanho da amostra é muito grande e os primeiros resultados são os mais importantes, a sensibilidade é bem menos utilizada para a avaliação de classificadores de especialistas em relação à precisão.

Já a métrica *R-Precision* (Rprec) é uma variação das duas acima, só que nesse caso ela é calculada levando em consideração o número R de especialistas da amostra. Uma propriedade interessante dessa métrica é que levando em conta os R primeiros resultados, o valor da precisão é igual ao valor da sensibilidade nesse ponto para classificadores binários.

A métrica *Average Precision@n* (AP@n) apresenta uma ideia geral da qualidade dos resultados, sendo calculada como a precisão média levando em conta a quantidade de acertos e a posição dos resultados relevantes entre os n primeiros. A fórmula do AP@n é apresentada na equação 6:

$$AP@n = \frac{\sum_{i \in n} P@i * R_{\delta}(i)}{n} \quad (6)$$

onde $R_{\delta}(i)$ é uma função que retorna 1 caso o resultado i seja relevante e 0 caso contrário.

Caso o objetivo seja indicar a precisão levando em conta todos os especialistas, essa métrica será denominada como *Average Precision* (AP), de modo que o n utilizado é a posição do último resultado relevante. A média da *Average Precision* para diversas amostras é chamada de *Mean Average Precision* (MAP).

Para conjuntos onde existam diversos níveis de especialidade podem ser empregadas métricas como a *Discounted Cumulative Gain* (DCG), que calcula a qualidade da classificação obtida de acordo com a pontuação de cada candidato, de modo que quanto melhor a ordem dos candidatos, maior a pontuação final. A fórmula do DCG é apresentada na equação 7:

$$DCG(X) = \sum_{i \in X} \frac{2^{p(i)} - 1}{\log_2(p(i) + 1)} \quad (7)$$

A variável X representa a lista testada e $p(i)$ a pontuação do i -ésimo classificado. O valor obtido para o DCG é geralmente normalizado em relação a pontuação da lista ideal. Essa métrica normalizada é denominada nDCG, que é apresentada na equação 8:

$$nDCG = \frac{DCG(O)}{DCG(I)} \quad (8)$$

Sendo O a lista obtida pela técnica de identificação de especialistas e I a lista ideal.

A Acurácia calcula a posição do melhor classificado na lista obtida. Seja r_1 o primeiro resultado da lista ideal e $pos(r_1, O)$ a posição de r_1 na lista obtida, o valor da acurácia é $ac = \frac{|O| - pos(r_1, O) + 1}{|O|}$, de modo que o valor diminui linearmente de acordo com a posição de r_1 na lista obtida.

3.4 Considerações sobre a revisão

A partir da revisão sistemática realizada, foi possível notar uma grande diversidade de abordagens para a identificação de especialistas nas redes sociais. Esta seção apresenta alguns dados referentes à pesquisa realizada.

A figura 4 apresenta o diagrama de Venn onde cada conjunto representa uma classe de estratégia, e cada interseção refere-se a uma categoria de estratégias mistas. Nela é possível notar que 22 dos 26 trabalhos que propuseram novas técnicas utilizam a estrutura de rede para a identificação de especialistas, devido ao escopo da pesquisa incluir redes sociais. As estratégias de perfil, por outro lado, foram aplicadas apenas 7 vezes.

Nas quatro abordagens onde não foram apresentados dados envolvendo métricas de rede para a identificação de especialistas, a estrutura foi usada indiretamente para a restrição do escopo de busca dos especialistas (ZHANG et al., 2012) ou foram extraídos dados de redes sociais virtuais para obtenção de seus dados pessoais e/ou de conhecimento (BOK et al., 2015; MOVSHOVITZ-ATTIAS et al., 2013; SRBA; GRZNAR; BIELIKOVA, 2015).

Entre os algoritmos que utilizam a estrutura de rede para a identificação de especialistas, 15 dos 22 aplicaram algoritmos de propagação com objetivo de aprimorar seus resultados. Já entre os 19 algoritmos listados por usarem o conhecimento dos candidatos para a identificação de especialistas, são recorrentes técnicas de extração de informações, como as detalhadas na seção 2.5, sendo essas aplicadas nove vezes.

Alguns dos artigos selecionados não produziram nenhuma técnica de identificação de especialistas, mas contribuíram revisando publicações da área (JOUNG et al., 2013; BONCHI et al., 2011; KHAN; WU; YAN, 2012), enquanto outros trabalhos apresentaram ferramentas e aplicações para a identificação de especialistas.

Figura 4 – Distribuição dos artigos de acordo com a estratégia

Fonte: Jamison José da Silva Lima, 2017

O trabalho realizado por Barbosa, Souza e Oliveira (2014) propõe a utilização de algoritmos de identificação de especialistas para aprimorar um processo manual de seleção de candidatos.

Heil, Wild e Gaedke (2014) apresentam uma ferramenta para identificação de especialistas em sistemas de gerenciamento.

O trabalho desenvolvido por Silva et al. (2012) realiza um experimento para tentar entender as motivações para os usuários buscarem por especialistas e a partir daí propor uma nova ferramenta para a identificação de especialistas a partir dos resultados obtidos por seu experimento.

O modelo proposto por Fazel-Zarandi et al. (2011) apresenta uma ferramenta que leva em consideração as motivações do usuário que está buscando os especialistas, de modo que para cada categoria de motivação é utilizado um algoritmo diferente. As categorias abrangem nível de especialidade, proximidade, homofilia, número de citações e popularidade.

A tabela 1 compila as categorias de informação empregadas pelos trabalhos incluídos nessa revisão, além do tipo de validação dos resultados em cada artigo revisado. As validações por meio de classificações manuais são realizadas por pessoas com conhecimento na área, geralmente por meio de questionários. Já as técnicas de validação que realizam a extração de informações referentes a especialidade dos candidatos aproveitam dados

coletados sobre os autores para inferir sua especialidade, sendo utilizados geralmente dados coletados de redes sociais e fóruns de resolução de perguntas.

Tabela 1 – Classes de estratégias e métodos para validação de dados usados em cada artigo

Artigo	Algoritmo			Validação	
	Perfil	Rede	Conhecimento	Manual	Extração
BARBOSA, DE SOUZA e OLIVEIRA 2014		Ferramenta			
BOK <i>et al.</i> 2015			X		X
BONCHI <i>et al.</i> 2011		Revisão			
BOZZON <i>et al.</i> 2013		X	X	X	
CAMPBELL <i>et al.</i> 2003		X	X	X	
CHAGAS, PÉREZ-ALCÁZAR E DIGIAMPIETRI 2015	X	X	X	X	
CHEN <i>et al.</i> 2011		X	X		X
DA SILVA <i>et al.</i> 2012		Ferramenta			
DENG <i>et al.</i> 2012		X	X	X	
FAZEL-ZARANDI <i>et al.</i> 2011		Ferramenta			X
HEIL, WILD e GAEDKE 2014		Ferramenta			
JOUNG <i>et al.</i> 2013		Revisão			X
KARDAN e MOHTAJ 2013		X			X
KARDAN, OMIDVAR E FARAHMANDNIA 2011		X		X	
KHAN, WU e YAN 2012		Revisão			
KLEINBERG 1999		Algoritmo de Propagação			
LI, N <i>et al.</i> 2012	X	X			
LI, X <i>et al.</i> 2013	X	X	X		X
LI, Y <i>et al.</i> 2012		X	X	X	
LIU e LI 2015	X	X	X	X	
LIU <i>et al.</i> 2013		X	X	X	
MOVSHOVITZ-ATTIAS <i>et al.</i> 2013			X		X
PADHYE, MUKHERJEE e SINHA 2014		X			
PAGE <i>et al.</i> 1999		Algoritmo de Propagação			
PUTTAN, e SENIVONGSE 2013		X	X	X	
SHAHRIARI, PAREKODI e KLAMMA 2015		X			X
SMIRNOVA 2011		X	X		X
SOUZA <i>et al.</i> 2013		X	X	X	
SRBA, GRZNAR e BIELIKOVA 2015	X		X		X
WELSER <i>et al.</i> 2011		X	X		
YENITERZI e CALLAN 2014		X	X	X	X
ZHANG <i>et al.</i> 2012	X				X
ZHANG, ACKERMAN e ADAMIC 2007		X		X	
ZHANG, TANG e LI 2007	X	X	X	X	
ZHAO <i>et al.</i> 2015		X	X		X
Total	7	22	19	13	13

Fonte: Jamison José da Silva Lima, 2017

Um dos grandes desafios da área é a comparação entre algoritmos diferentes, pela falta de bases de dados abertas de amostragem e avaliação, o que facilitaria a comparação entre técnicas diferentes, e também pela falta de generalização das abordagens, que por levarem em conta fontes de dados diferentes, acabam tornando-se dependentes dos dados disponíveis em seu contexto.

Neste trabalho foi realizada a comparação entre o algoritmo proposto com a abordagem proposta por Chagas, Perez-Alcazar e Digiampietri (2015), devido a utilização da mesma estrutura e do mesmo conjunto de dados pelos dois algoritmos.

4 Solução proposta

Este capítulo apresenta a solução proposta nesse projeto, detalhando as informações coletadas, a estrutura utilizada e os algoritmos implementados. Este projeto estendeu o trabalho produzido por Chagas, Perez-Alcazar e Digiampietri (2015), buscando melhorar a precisão dos resultados, ao incluir mudanças e restrições nos algoritmos propostos.

4.1 Informações e estruturas utilizadas

Figura 5 – Diagrama de dados do algoritmo desenvolvido

Fonte: Jamison José da Silva Lima, 2017

Foram obtidas as seguintes informações de cada autor para a identificação de especialistas: identificador Lattes, lista de coautores e número de publicações em coautoria entre cada par de coautores, áreas de atuação, títulos das publicações e os conceitos Qualis dos veículos que cada autor publicou.

Com relação à estrutura, foi gerada uma rede social de coautorias com atributos em seus nós e arestas, na qual cada nó representa um pesquisador, possuindo atributos indicando seu identificador, seus coautores, uma lista contendo os títulos de suas publicações, uma lista das áreas de atuação declaradas pelo pesquisador e os valores referentes aos

Scores Alfa, Beta e Propagated. Já cada aresta, que representa uma relação de coautoria, possui como atributos o par de coautores e o número de publicações em coautoria entre eles.

A figura 5 apresenta um diagrama com as informações de entrada e saída aplicadas nesse projeto. No diagrama é possível notar que a partir dos bancos de dados Lattes e Qualis-Periódicos é formado o banco de dados usado nesse projeto (DIGIAMPIETRI et al., 2014), e a partir dele são obtidos dados sobre os autores, suas publicações e os estratos Qualis das revistas em que foram publicados os artigos e as coautorias.

4.2 Algoritmos implementados

Nessa seção o algoritmo proposto neste projeto é detalhado, de modo a explicar as decisões tomadas em sua implementação. O algoritmo 1 apresenta as principais etapas para a identificação de especialistas.

Algoritmo 1 Estrutura básica do algoritmo de identificação de especialistas proposto

```

1: procedure CALCULAESPECIALISTAS()
2:   INICIALIZA GRAFO(Arquivo arestas)
3:   INICIALIZA AUTORES(Arquivo publicacoes)
4:   INICIALIZA AREAS(Arquivo areas)
5:   INICIALIZA LISTA PALAVRAS(String area, int n, int maxAreas)
6:   CALCULA SCORE ALFA(String area, int n, int maxAreas)
7:   CALCULA SCORE BETA(String areaQualis)
8:   PREPARA PROPAGATED(String area, bool subrede, bool somaAlfaBeta)
9:   CALCULA PROPAGATED(String area, double d, double epsilon, bool removeOrfao)

```

Fonte: Jamison José da Silva Lima, 2017

A inicialização do grafo é realizada por meio de um arquivo listando as relações de coautoria e o número de coautorias entre cada par de pesquisadores. A inicialização dos autores é realizada para que cada autor tenha uma lista contendo os títulos de suas publicações, e logo após é inicializada a lista de áreas que o pesquisador declara pesquisar. Todos os dados passados como entrada para esse algoritmo foram coletados do banco de dados gerado por Digiampietri et al. (2014), como já descrito na seção 2.4. Então são calculados os Scores Alfa e Beta pelo cálculo do TF-IDF a partir da lista de palavras de cada pesquisador e pelo Qualis da lista de publicações de cada autor. O Score Propagated realiza a propagação dos valores calculados anteriormente para cada autor a partir do

algoritmo proposto por Zhang, Tang e Li (2007). Nas próximas subseções serão detalhados todos os algoritmos implementados.

4.2.1 Inicialização do grafo

O algoritmo 2 apresenta a inicialização do grafo, que é realizada pela leitura de um arquivo de texto em que cada linha contém um par de *ids* e um peso, de modo que cada *id* representa o identificador de um pesquisador cadastrado no Lattes e o peso indica o número de coautorias entre eles.

Algoritmo 2 Inicialização do grafo

```

1: procedure INICIALIZA_GRAFO(Arquivo arestas)
2:   while arestas tiver próxima linha contendo “id1 id2 peso” do
3:     if id1 não foi adicionado no grafo then
4:       adiciona o nó id1 no grafo
5:     if id2 não foi adicionado no grafo then
6:       adiciona o nó id2 no grafo
7:     adiciona (id1, id2, peso) na lista de arestas do grafo

```

Fonte: Jamison José da Silva Lima, 2017

A estrutura do grafo possui uma lista de nós (*listaAutores*) e uma lista de arestas (*listaArestas*), que são necessárias para o funcionamento do algoritmo.

4.2.2 Inicialização das informações dos pesquisadores

Após a criação do grafo é realizada a leitura de um arquivo contendo os títulos das publicações de cada autor. Na solução proposta, diferentemente do algoritmo proposto por Chagas, Perez-Alcazar e Digiampietri (2015), onde os autores são inicializados somente para a obtenção do Score Propagated, cada autor possui uma lista de palavras, como mostrado no algoritmo 3.

Nesse algoritmo é realizada a remoção de *stopwords* do título, que são um conjunto de palavras largamente usadas nos títulos das publicações e irrelevantes ao contexto, incluindo alguns artigos, advérbios, preposições, verbos e pronomes. Além da aplicação de listas de *stopwords* já existentes, foram também removidos termos comuns presentes nos títulos das publicações coletadas para esse trabalho.

Algoritmo 3 Inicialização dos dados das publicações para cada autor

```

1: procedure INICIALIZAAUTORES(Arquivo publicacoes)
2:   while publicacoes tiver próxima linha contendo as variáveis titulo, idioma e id do
3:     if idioma for igual a “Português” ou “Inglês” then
4:       remove as stopwords e caracteres especiais do titulo
5:       executa o stemmer para o idioma
6:       adiciona as palavras stemizadas a lista de palavras do id

```

Fonte: Jamison José da Silva Lima, 2017

Também foram empregadas ferramentas de *stemização* (*stemming*) para os idiomas português e inglês disponíveis na biblioteca Apache Lucene¹. A stemização consiste na remoção dos sufixos de palavras derivadas, com o objetivo de agrupá-las de acordo com suas raízes. Foram incluídos apenas artigos em inglês e português pois 97,89% dos 8,56 milhões de publicações cadastradas na Plataforma Lattes foram escritas nesses idiomas.

Devido a grande quantidade de palavras nos subconjuntos de dados coletados, foi utilizada uma estrutura de *hash* contendo a versão stemizada de cada palavra encontrada para evitar a repetição dos algoritmos de stemização.

Após a inicialização dos documentos é realizada a leitura de um arquivo contendo a lista das áreas de atuação dos pesquisadores, como demonstrado no algoritmo 4.

Algoritmo 4 Inicialização da lista de áreas de atuação dos pesquisadores

```

1: procedure INICIALIZAAREAS(Arquivo areas)
2:   while areas tiver próxima linha contendo as variáveis id, subarea do
3:     adiciona subarea à lista de áreas de id

```

Fonte: Jamison José da Silva Lima, 2017

4.2.3 Cálculo do Score Alfa

Após a inicialização do grafo e da lista de palavras de cada pesquisador, é possível calcular a relevância das publicações de cada autor para a área por meio da métrica TF-IDF, obtendo assim o Score Alfa.

Como apresentado na seção 2.5, o TF-IDF é uma métrica estatística que calcula a relevância de um documento no corpus a partir da frequência e raridade de seus termos relevantes em relação ao corpus. Portanto, para a aplicação do TF-IDF é necessária a obtenção de uma lista de palavras relevantes para o tópico. Algumas abordagens atuais

¹ lucene.apache.org

aplicam ontologias de domínio, porém, devido aos desafios para encontrá-las ou gerá-las, neste trabalho foram empregadas listas de palavras relevantes para cada área selecionada.

O algoritmo 5 apresenta como foi gerada a lista das palavras utilizadas no cálculo do TF-IDF.

Algoritmo 5 Criação da lista de palavras para uma área

```

1: procedure INICIALIZALISTAPALAVRAS(String area, int n, int maxAreas)
2:   for all autor em listaAutores do
3:     if O autor declara publicar na área buscada e em até maxAreas áreas diferentes
       then
4:       adiciona todas as suas palavras na lista de candidatas
5:       aumenta a frequência de cada palavra na lista de candidatas em 1
6:   Return lista contendo as n palavras mais usadas entre as candidatas

```

Fonte: Jamison José da Silva Lima, 2017

O algoritmo de criação da lista de palavras seleciona todos os termos dos títulos de publicações envolvendo os autores que declaram publicar na área buscada e em até *maxAreas* áreas. Esse limite foi definido para controlar a quantidade de termos de outras áreas incluídos na lista de palavras. Com todos os termos candidatos selecionados é feito o cálculo da frequência dessas palavras no corpus, sendo então selecionados os *n* termos com maior frequência na lista de palavras candidatas. Esse número de termos é um parâmetro arbitrário dado como entrada para o algoritmo, que será otimizado no capítulo 5.

Com a lista de palavras relevantes selecionadas, é então calculado Score Alfa de cada autor, como apresentado no algoritmo 6.

Algoritmo 6 Cálculo do Score Alfa

```

1: procedure CALCULASCOREALFA(String area, int n, int maxAreas)
2:   listaPalavras = INICIALIZALISTAPALAVRAS(area, n, maxAreas)
3:   for all autor em listaAutores do
4:     autor.alfa = CALCULATFIDF(autor.listaPalavras, listaPalavras, listaAutores.size)
5:   normaliza alfa de todos autores no intervalo [0,1]

```

Fonte: Jamison José da Silva Lima, 2017

Na abordagem proposta por Chagas, Perez-Alcazar e Digiampietri (2015) cada publicação contém sua lista de palavras, portanto o Score Alfa de seu algoritmo calcula a relevância de cada revista científica em relação à área de acordo com seus títulos, dada uma ontologia selecionada previamente. Por outro lado, para esse trabalho o TF-IDF é calculado para cada autor, sendo utilizada a lista de palavras mais frequentes da área ao

invés de uma ontologia. Desse modo o TF-IDF de cada autor é igual a soma do TF-IDF para cada um dos termos relevantes contidos em sua lista de palavras.

4.2.4 Cálculo do Score Beta

O cálculo do Score Beta é realizado de acordo com os valores Qualis das publicações para a área onde estão sendo buscados os especialistas. Inicialmente foram obtidos os estratos Qualis de todas as publicações para todas as áreas por meio da plataforma Sucupira² para o quadriênio 2013-2016. A tabela 2 apresenta como exemplo os estratos Qualis para a revista *Computers in Biology and Medicine*³, que publica artigos envolvendo aplicações computacionais para as áreas de biociências e medicina.

Tabela 2 – Qualis da revista *Computers in Biology and Medicine*

Estrato Qualis	Comitê de Área
A2	ENGENHARIAS IV
B1	BIODIVERSIDADE
B1	ENGENHARIAS II
B1	MEDICINA III
B1	INTERDISCIPLINAR
B1	CIÊNCIA DA COMPUTAÇÃO
B1	MEDICINA II
B1	PSICOLOGIA
B2	MEDICINA I
B3	EDUCAÇÃO FÍSICA
B3	CIÊNCIAS BIOLÓGICAS II
B3	QUÍMICA
B3	BIOTECNOLOGIA
B3	CIÊNCIAS BIOLÓGICAS I
B5	ASTRONOMIA / FÍSICA

Fonte: Plataforma Sucupira, 2017

Esse exemplo ilustra o fato de que uma revista pode possuir estratos em áreas diferentes, pois a classificação é realizada considerando o escopo da área e se houveram publicações recentes realizadas por membros dos programas de pós-graduação da área, além da utilização de métricas como o fator de impacto.

Outro detalhe relevante é que os comitês de área Qualis são divididos de modo que alguns deles englobam diversas áreas relacionadas. Por exemplo, a área “Engenharias

² sucupira.capes.gov.br

³ www.journals.elsevier.com/computers-in-biology-and-medicine

IV” inclui entre seus programas de pós-graduação Engenharia Biomédica e Engenharia de Computação, áreas diretamente relacionadas ao conteúdo da revista.

Para este trabalho cada estrato recebeu um peso de modo que pesquisadores com diversas publicações em revistas importantes possuam melhores classificações na identificação de especialistas. A tabela 3 apresenta os pesos atribuídos a cada estrato, que são os mesmos adotados pelo comitê de área de Ciência da Computação.

Tabela 3 – Pesos atribuídos aos estratos Qualis de cada publicação

A1	A2	B1	B2	B3	B4	B5	C
100	85	70	50	20	10	5	0

Fonte: Jamison José da Silva Lima, 2017

Com os estratos coletados e com os pesos definidos, torna-se possível classificar os autores de acordo com suas publicações, de modo que cada autor recebe uma pontuação para cada área, definida como a soma das pontuações de seus artigos publicados para a área. Essas pontuações foram calculadas *a priori* e armazenadas em arquivos. A Equação 9 apresenta a normalização do Score Beta em relação às maiores pontuações Qualis da área e do autor.

$$Beta(id, area) = \frac{qualis(id, area)}{qualisMax(area)} * \frac{qualis(id, area)}{qualisMax(id)} \quad (9)$$

A função $qualis(id, area)$ representa a pontuação total do Qualis do autor id para a área buscada, $qualisMax(area)$ e $qualisMax(id)$ representam a maior pontuação Qualis para a área buscada e para o autor respectivamente.

Além da fórmula composta apresentada, podem ser selecionadas como fator de normalização o valor máximo da área pesquisada ($\frac{qualis(id, area)}{qualisMax(area)}$) ou também o Qualis máximo do autor ($\frac{qualis(id, area)}{qualisMax(id)}$).

4.2.5 Cálculo do Score Propagated

O Score Propagated é calculado por meio de um algoritmo de propagação semelhante ao proposto por Chagas, Perez-Alcazar e Digiampietri (2015). Antes da execução são realizados os seguintes passos preliminares:

- É obtida a média dos Scores Alfa e Beta, que é normalizada e utilizada como valor inicial do algoritmo de propagação.

- Cada nó recebe sua lista de arestas adjacentes e também é armazenada a soma dos pesos de suas relações, que são necessárias para o cálculo do Score Propagated.
- No caso de redes contendo diversas áreas, podem ser removidas as arestas não relacionadas a área pesquisada a partir da seguinte regra: se nenhum dos autores declarou que pertence a área em que os especialistas estejam sendo buscados, a relação é desfeita por não pertencer a área em que os especialistas estão sendo buscados.

O algoritmo 7 apresenta a abordagem utilizada para a propagação dos resultados.

Algoritmo 7 Cálculo do Score Propagated

```

1: procedure CALCULAPROPAGATED(double d, double epsilon, bool removeOrfao)
2:   repeat
3:     for all autor em listaAutores do
4:       if  $V_{autor} \neq \emptyset$  then
5:          $autor.rankTemp = autor.rank * (1 - d) + d * \sum_{v \in V_{autor}} (v.rank * \frac{peso(autor,v)}{pesoTotal_v})$ 
6:          $erroMedio = erroMedio + \frac{|autor.rankTemp - autor.rank|}{listaAutores.size}$ 
7:       else if removeOrfao then autor.rankTemp = 0
8:     for all autor em listaAutores do
9:       autor.rank = autor.rankTemp
10:    NORMALIZAPROPAGATED()           ▷ normaliza os Scores no intervalo [0,1]
11:  until erroMedio > epsilon

```

Fonte: Jamison José da Silva Lima, 2017

Nele, d é o fator de propagação, passado como parâmetro para limitar a proporção do Score que é repassado aos vizinhos, $epsilon$ é o erro médio mínimo para o algoritmo continuar sendo executado, V_{autor} é a lista de vizinhos de *autor*, *autor.rank* é o valor do Score Propagated, *autor.rankTemp* é o Score Propagated temporário repassado para *autor.rank* no final de cada iteração, $peso(autor,v)$ representa o peso da relação entre *autor* e seu vizinho v , $pesoTotal_v$ representa a soma dos pesos de todas as relações do nó v . Após sua execução, a variável *rank* de cada autor representa seu nível de especialidade. Ordenando os resultados do maior para o menor *rank* é obtida a lista de especialistas.

4.3 Considerações

Esse capítulo apresentou os dados coletados, a estrutura usada e o algoritmo proposto para a identificação de especialistas, consistindo de três etapas de classificação dos pesquisadores: a primeira envolvendo a relevância dos títulos de suas publicações, a

segunda etapa calculando a qualidade de suas publicações para a área e na terceira etapa é calculada a influência de cada pesquisador na rede de coautorias a partir do algoritmo de propagação proposto.

As principais variações propostas nesse projeto em relação ao algoritmo idealizado por Chagas, Perez-Alcazar e Digiampietri (2015) são as seguintes:

- Cálculo dos Scores Alfa e Beta para os autores ao invés das publicações, facilitando a análise e o aprimoramento dos resultados parciais.
- Remoção de palavras recorrentes além das listas básicas de stopwords.
- Geração automática de uma ontologia simplificada ao invés da utilização de ontologias existentes, que são difíceis de serem encontradas.
- Cálculo do Score Beta normalizado a partir da pontuação do autor para todas as áreas.
- Generalização do algoritmo Score Propagated, não sendo necessária a remoção dos pesquisadores que não publicam na área buscada.

No próximo capítulo são apresentados os resultados obtidos a partir dos conjuntos de dados coletados e do algoritmo implementado.

5 Resultados

Nesse capítulo são apresentados os resultados obtidos pelo algoritmo proposto nesse projeto após a otimização de seus parâmetros e a comparação dos resultados em relação ao algoritmo base (CHAGAS; PEREZ-ALCAZAR; DIGIAMPIETRI, 2015). As comparações foram realizadas a partir de três conjuntos de pesquisadores extraídos da Plataforma Lattes.

Este capítulo está organizado da seguinte maneira. Primeiramente são apresentadas as informações iniciais na seção 5.1. Na seção 5.2 são apresentados os resultados parciais e os parâmetros da abordagem proposta são otimizados. Na seção 5.3 são apresentados os resultados obtidos pela abordagem proposta em comparação ao algoritmo base. Na seção 5.4 são apresentadas as considerações sobre os resultados.

5.1 Dados coletados

Para esse trabalho foram coletados dados de três subconjuntos de pesquisadores cadastrados na Plataforma Lattes, conforme descrito na seção 1.2.

O primeiro conjunto de pesquisadores é composto de 13.357 doutores que recebiam a Bolsa Produtividade em Pesquisa do CNPq em 2016. O segundo conjunto contém 3.343 pesquisadores que declararam atuar em Inteligência Artificial em seus currículos Lattes. O conjunto de Nanotecnologia incluiu 912 pesquisadores selecionados de acordo com a metodologia utilizada por Chagas, Perez-Alcazar e Digiampietri (2015).

5.2 Otimização dos parâmetros e resultados preliminares

Nesta seção são apresentados os resultados parciais e a otimização de cada parâmetro a partir dos resultados para três áreas do conjunto de bolsistas produtividade: Ciência da Computação, Química e Física. Essas três áreas foram selecionadas arbitrariamente.

Para a seleção dos valores iniciais dos parâmetros foram realizados testes que apontaram a direção dos melhores resultados para a área da Ciência da Computação. A tabela 4 apresenta os valores iniciais dos parâmetros utilizados.

Tabela 4 – Valores padrão dos parâmetros utilizados nos testes preliminares

Score	Nome da variável	Descrição	Valor
Alfa	<i>areaMaxSize</i>	Número máximo de áreas para adicionar na ontologia	<i>1</i>
	<i>ontologiaSize</i>	Tamanho mínimo da ontologia	<i>300</i>
Beta	<i>maxArea</i>	Normalizar a pontuação Qualis levando em consideração o valor máximo para a área	<i>true</i>
	<i>maxAutor</i>	Normalizar a pontuação Qualis levando em consideração o valor máximo para o autor	<i>true</i>
Propagated	<i>somaAlfaBeta</i>	Se verdadeiro, calcula a média aritmética de Alfa e Beta, senão é calculada a média geométrica	<i>false</i>
	<i>subrede</i>	Remove arestas entre autores que não publicaram na área	<i>true</i>
	<i>removeOrfao</i>	Remove nós de autores que não possuem nenhuma aresta	<i>false</i>
	<i>d</i>	Taxa de propagação no intervalo (0,1)	<i>0,05</i>
	<i>epsilon</i>	Erro médio utilizado como fator de parada	<i>0,01</i>

Fonte: Jamison José da Silva Lima, 2017

As próximas subseções descrevem a otimização de cada um desses, iniciando pelos envolvidos no Score Alfa na próxima subseção, sendo depois descritas as otimizações dos Scores Beta e Propagated nas subseções seguintes. A abordagem utilizada para a otimização foi incremental, de modo que os parâmetros do Score Alfa foram otimizados primeiro, e em seguida os Scores Beta e Propagated tiveram seus parâmetros otimizados, devido a imensa quantidade de combinações possíveis entre os resultados.

Para a avaliação foram selecionadas métricas envolvendo a precisão dos resultados, sendo elas *R-Precision*, *Average Precision* e *Precision@n*. Ao final do treinamento foram realizados testes envolvendo a métrica MAP para 20 áreas do conjunto de bolsistas produtividade, de modo que o melhor resultado foi selecionado para realização dos testes finais para os três conjuntos de dados coletados.

5.2.1 Score Alfa

Nesta seção são apresentados os resultados obtidos para o conjunto de bolsistas envolvendo as variações nos parâmetros do Score Alfa, que é calculado para cada autor a

partir dos títulos de seus artigos, de modo que os autores com o maior número de publicações envolvendo tópicos que sejam identificados como relevantes para a área possuem maior valor.

A tabela 5 apresenta os resultados obtidos para a identificação de especialistas envolvendo apenas o Score Alfa para o conjunto da Ciência da Computação, sem os cálculos dos Scores Beta e Propagated, sendo assim possível se ter uma ideia da qualidade dessa técnica isoladamente para a identificação de especialistas.

Tabela 5 – Resultados da identificação de especialistas em Ciência da Computação para o Score Alfa

Precision@n		Recall@n	
P@1	0	R@1	0
P@5	0,4	R@5	0,0155
P@10	0,4	R@10	0,03101
P@25	0,28	R@25	0,05426
P@50	0,34	R@50	0,13178
P@100	0,3	R@100	0,23256
P@250	0,252	R@250	0,48837
P@500	0,196	R@500	0,75969
R-precision:	0,295	No. Experts:	129

Fonte: Jamison José da Silva Lima, 2017

A partir dos resultados, observa-se que apenas quatro pesquisadores que recebem Bolsa Produtividade nível 1 foram encontrados entre os 10 primeiros, e que menos de um terço dos 129 Bolsistas Produtividade nível 1 da área foram classificados entre os 129 primeiros (como mostrado pelo resultado da métrica R-Precision).

A seguir são discutidos os resultados a partir da otimização dos parâmetros do Score Alfa, sendo eles o número máximo de áreas e o tamanho da lista de palavras, os quais serão detalhados a seguir. Todos os próximos resultados apresentados nessa subseção levarão em consideração a execução completa do algoritmo, mantendo constantes os parâmetros envolvendo os Scores Beta e Propagated de acordo com os valores apresentados na tabela 4.

Número máximo de áreas da lista de palavras

O número de áreas declaradas pelo autor em seu currículo Lattes é utilizado como parâmetro durante a obtenção da lista de palavras. Caso configurado para o valor mínimo (uma área), serão utilizados os títulos das publicações apenas dos autores que declararam atuar em uma única área de atuação/interesse (no caso, a área na qual os especialistas

estão sendo procurados). Se o número máximo permitido aumentar para duas áreas, são utilizados os títulos das publicações de autores que declararam atuar em duas áreas diferentes, sendo uma delas a área na qual os especialistas estão sendo procurados e assim por diante. Os termos utilizados nos títulos das publicações são adicionados aos candidatos da lista de palavras, de modo que os termos mais recorrentes entre os selecionados serão utilizados no cálculo do Score Alfa.

Os resultados dos testes envolvendo esse parâmetro são mostrados na tabela 6. Para cada variação testada (valores 1, 2 e 3) foi calculada sua precisão média a partir de testes envolvendo as três áreas selecionadas, de modo que para cada uma delas foram testados três tamanhos diferentes para a lista de palavras (300, 500 e 700).

Tabela 6 – Precisão média em relação ao número máximo de áreas permitidas para inclusão de termos na lista de palavras

Precisão	Número máximo de áreas		
	1	2	3
R-Precision	0,4997	0,4977	0,4901
P@1	1,0000	1,0000	1,0000
P@5	0,6889	0,6444	0,6444
P@10	0,6000	0,5556	0,5556
P@25	0,6089	0,5911	0,5822
P@50	0,5933	0,5978	0,5978
P@100	0,5556	0,5378	0,5322
P@250	0,4662	0,4653	0,4569
P@500	0,3671	0,3578	0,3531

Fonte: Jamison José da Silva Lima, 2017

Os resultados mostram que a utilização dos termos de pesquisadores que atuam apenas na área buscada para a criação da lista de palavras apresentam a melhor precisão. Existem dois fatores principais que podem explicar estes resultados: a adição de títulos de artigos pertencentes a áreas diferentes pode ocorrer caso o pesquisador publique em outras áreas além da área pesquisada e, na abordagem utilizada, nenhuma técnica foi incluída para a identificação das áreas de cada artigo, assim, foram utilizadas apenas as áreas indicadas pelos autores.

Tamanho mínimo da lista de palavras

Com o número de áreas selecionado, foram realizados testes para avaliar o desempenho de acordo com o tamanho mínimo da lista de palavras, que é denominado assim pois após a realização do cálculo da frequência dos termos selecionados como candidatos, é

feita a inclusão dos *tamListaPalavras* termos mais utilizados pelos pesquisadores da área, sendo que a frequência do último termo é utilizada como limiar, sendo assim incluídos termos além do total caso possuam a mesma frequência. Como cada área possui uma lista de publicações diferente, há uma dificuldade inerente em selecionar esse número.

Para selecionar o tamanho ideal foi decidido utilizar a precisão dos primeiros resultados, que podem ser considerados os mais importantes para algoritmos de busca (SU; YUAN; ZHU, 2015). As figuras 6a e 6b mostram a variação da métrica *Average Precision* em relação ao tamanho da lista de palavras para as áreas selecionadas, com os valores 10 e 25 selecionados como candidatos para n .

Figura 6 – Resultados das métricas AP@10 e AP@25 dada a variação do tamanho da lista de palavras

Fonte: Jamison José da Silva Lima, 2017

Para a métrica AP@10, a área de Ciência da Computação obteve o resultado 0,866 para os tamanhos 60, 651, 684 e 731. Para a área de Química, o melhor resultado foi 0,725 com 17 palavras. Na área da Física o maior *Average Precision* entre os 10 primeiros resultados foi de 0,493 para a lista contendo entre 54 e 61 palavras.

Já para a métrica AP@25 a área da Computação teve o melhor resultado com 415 e 437 palavras na lista e valor 0,715, a Química com tamanho 184 e valor 0,505 e 0,363 para a área de Física com 74 palavras.

Para os próximos testes, serão utilizados os resultados da métrica *Average Precision* com n igual a 25 devido a menor variância nesses testes.

Além do uso da precisão média, foi decidida a utilização do valor da métrica R-Precision, que representa a precisão em um ponto intermediário dos resultados. A figura 7 apresenta os resultados obtidos para a métrica *R-Precision* de acordo com o tamanho da lista.

Figura 7 – Variação do R-Precision de acordo com o tamanho da lista de palavras

Fonte: Jamison José da Silva Lima, 2017

Os tamanhos da lista de palavras que maximizam o valor de *R-Precision* de cada área são: 215 e 300 para a Ciência da Computação com *R-Precision* igual a 0,527, a lista com 596 e 622 palavras apresentaram o valor a 0,528 para a área da Química. Já para a Física 2302 termos na lista obteve o melhor valor de *R-Precision*, igual 0,577.

Como é possível reparar a partir desses resultados, a precisão média e o R-Precision separadamente não são suficientes para identificar quais seriam os tamanhos ideais. A fim de identificar o tamanho da lista de palavras a ser utilizado nos demais testes, optou-se por otimizar a média geométrica das métricas *AP@25* e *R-Precision*. A partir desses resultados foram selecionados os seguintes tamanhos para cada uma das áreas: 300 para a Ciência da Computação, 184 para a Química e 141 para a Física. Como pôde ser notado a partir dos tamanhos selecionados, há uma dificuldade inerente em fixar o tamanho da lista pois cada área possui uma lista de publicações diferente.

A tabela 7 apresenta a precisão e sensibilidade dos resultados após a otimização dos parâmetros Alfa.

Tabela 7 – Precisão e sensibilidade para os melhores tamanhos da lista de palavras

Valores	Áreas		
	Computação	Química	Física
<i>Tam. ontologia</i>	300	184	141
<i>No. Especialistas</i>	129	286	435
<i>R-Precision</i>	0,527	0,493	0,469
<i>P@1</i>	1	1	1
<i>P@5</i>	0,8	0,8	0,6
<i>P@10</i>	0,8	0,8	0,5
<i>P@25</i>	0,8	0,64	0,56
<i>P@50</i>	0,64	0,6	0,46
<i>P@100</i>	0,54	0,58	0,52
<i>P@250</i>	0,36	0,532	0,476
<i>P@500</i>	0,228	0,38	0,456
<i>R@1</i>	0,008	0,004	0,002
<i>R@5</i>	0,031	0,014	0,007
<i>R@10</i>	0,062	0,028	0,011
<i>R@25</i>	0,155	0,056	0,032
<i>R@50</i>	0,248	0,105	0,053
<i>R@100</i>	0,419	0,203	0,120
<i>R@250</i>	0,698	0,465	0,274
<i>R@500</i>	0,884	0,664	0,524

Fonte: Jamison José da Silva Lima, 2017

Comparando os resultados apresentados na tabela 7 com os obtidos pelo cálculo do Score Alfa (tabela 5) que possui os mesmos parâmetros utilizados, é possível notar que há uma melhora significativa nos resultados obtidos, com destaque principal para os 25 primeiros classificados, sendo 80% deles especialistas para a execução completa contra 28% encontrados apenas pelo Score Alfa.

5.2.2 Score Beta

Nesta seção são apresentados os resultados obtidos para o conjunto de treinamento envolvendo as variações nos parâmetros do Score Beta. O Score Beta é calculado para cada autor a partir da soma dos estratos Qualis de suas publicações para a área pesquisada, de acordo com os valores mostrados na tabela 3, de modo que os autores com mais publicações em revistas de maior qualidade para a área terão os melhores resultados.

A tabela 8 apresenta os valores dos parâmetros utilizados nos testes dessa seção. Como o Score Alfa já foi otimizado, os conjuntos de parâmetros utilizados nos testes

dessa seção possuirão valores diferentes para a variável *tamListaPalavras*. Desse modo, os parâmetros Beta serão otimizados separadamente para cada uma das áreas.

Tabela 8 – Valores padrão dos parâmetros utilizados nos testes do Score Beta

Score	Nome da variável	Descrição	Computação	Química	Física
Alfa	<i>areaMaxSize</i>	Número máximo de áreas para adicionar na ontologia	1		
	<i>ontologiaSize</i>	Tamanho da ontologia	300	184	141
Beta	<i>maxArea</i>	Normalizar a pontuação Qualis levando em consideração o valor máximo da área	true		
	<i>maxAutor</i>	Normalizar a pontuação Qualis levando em consideração o valor máximo do autor	true		
Propagated	<i>somaAlfaBeta</i>	Se verdadeiro, calcula a média aritmética de Alfa e Beta, senão é calculada a média geométrica	false		
	<i>subrede</i>	Remove arestas entre autores que não publicaram na área	true		
	<i>removeOrfao</i>	Remove nós de autores que não possuem nenhuma aresta	false		
	<i>d</i>	Taxa de propagação no intervalo (0,1)	0,05		
	<i>epsilon</i>	Erro médio utilizado como fator de parada	0,01		

Fonte: Jamison José da Silva Lima, 2017

A tabela 9a apresenta os resultados de precisão e revocação envolvendo apenas o Score Beta para a área da Ciência da Computação. Em comparação aos resultados obtidos para o Score Alfa (tabela 5), o Score Beta apresenta melhores resultados para todos os pontos de precisão apresentados. Já a média dos Scores Alfa e Beta apresentada na tabela 9b obteve resultados ainda melhores, mas inferiores à execução completa apresentada anteriormente na tabela 7.

Tabela 9 – Resultados parciais da identificação de especialistas em Ciência da Computação envolvendo o Score Beta

(a) Beta

Precision@n	Recall@n
P@1 0	R@1 0
P@5 0,6	R@5 0,02326
P@10 0,7	R@10 0,05426
P@25 0,56	R@25 0,10853
P@50 0,52	R@50 0,20155
P@100 0,48	R@100 0,37209
P@250 0,324	R@250 0,62791
P@500 0,208	R@500 0,8062
R-precision: 0,44186	No. Experts: 129

(b) Alfa + Beta

Precision@n	Recall@n
P@1 0	R@1 0
P@5 0,6	R@5 0,02326
P@10 0,8	R@10 0,06202
P@25 0,64	R@25 0,12403
P@50 0,52	R@50 0,20155
P@100 0,48	R@100 0,37209
P@250 0,348	R@250 0,67442
P@500 0,226	R@500 0,87597
R-precision: 0,44961	No. Experts: 129

Fonte: Jamison José da Silva Lima, 2017

Esses resultados reforçam a hipótese de que utilizar abordagens compostas para a identificação de especialistas apresenta melhores resultados do que a utilização em isolado das abordagens.

Normalização da pontuação Qualis

O Score Beta possui dois parâmetros binários, *maxArea* e *maxAutor*. O primeiro deles envolve a normalização em relação a pontuação Qualis máxima da área, já o segundo representa a normalização das pontuações de cada autor a partir de sua área com pontuação máxima. Caso as duas normalizações sejam selecionadas simultaneamente é calculado o produto delas e denominada como normalização composta, como mostrado na Equação 9. Caso nenhuma das normalizações seja selecionada, é escolhida por padrão a normalização pelo máximo da área.

Na figura 8 são mostrados os resultados das métricas $AP@25$ e $R\text{-Precision}$ para cada área selecionada de acordo com as três variações de normalização para as pontuações Qualis.

Figura 8 – Resultados para as métricas R-Precision e $AP@25$ dada a variação dos parâmetros do Score Beta

Fonte: Jamison José da Silva Lima, 2017

Os resultados obtidos indicam que para a área de Ciência da Computação e para a Química, os melhores resultados para a identificação de especialistas envolvem a

normalização composta. Já para a área de Física, a normalização por autor foi selecionada, sendo relevante notar que a abordagem composta obteve os piores resultados.

5.2.3 Score Propagated

Por fim, foi realizada a otimização dos parâmetros do Score Propagated, que consiste na utilização dos Scores Alfa e Beta obtidos anteriormente como entrada para o algoritmo de propagação proposto, levando em consideração os relacionamentos de coautoria da rede.

A tabela 10 apresenta os valores que serão utilizados inicialmente nos experimentos desta seção para cada uma das áreas testadas, de acordo com os parâmetros selecionados anteriormente.

Tabela 10 – Valores padrão dos parâmetros utilizados nos testes do Score Propagated

Score	Nome da variável	Descrição	Computação	Química	Física
Alfa	<i>areaMaxSize</i>	Número máximo de áreas para adicionar na ontologia	<i>1</i>		
	<i>ontologiaSize</i>	Tamanho da ontologia	<i>300</i>	<i>184</i>	<i>141</i>
Beta	<i>maxArea</i>	Normalizar a pontuação Qualis levando em consideração o valor máximo da área	<i>true</i>	<i>true</i>	<i>false</i>
	<i>maxAutor</i>	Normalizar a pontuação Qualis levando em consideração o valor máximo do autor	<i>true</i>		
Propagated	<i>somaAlfaBeta</i>	Se verdadeiro, calcula a média aritmética de Alfa e Beta, senão é calculada a média geométrica	<i>false</i>		
	<i>subrede</i>	Remove arestas entre autores que não publicaram na área	<i>true</i>		
	<i>removeOrfao</i>	Remove nós de autores que não possuem nenhuma aresta	<i>false</i>		
	<i>d</i>	Taxa de propagação no intervalo (0,1)	<i>0,05</i>		
	<i>epsilon</i>	Erro médio utilizado como fator de parada	<i>0,01</i>		

Fonte: Jamison José da Silva Lima, 2017

Na tabela 11 são apresentados os resultados parciais envolvendo o Score Propagated para a identificação de especialistas da Ciência da Computação. Nos resultados obtidos para a execução solo do Score Propagated na tabela 11a, cada candidato recebe o mesmo valor inicial antes da execução do algoritmo de propagação, já nas outras execuções mostradas nas tabelas 11b e 11c, os Scores Alfa ou Beta são passados como entrada para o algoritmo de propagação. É possível observar que os melhores resultados parciais apresentados envolvem a composição dos Scores Beta e Propagated, entretanto não são bons o suficiente para superar os resultados envolvendo a média dos Scores Alfa e Beta apresentados na tabela 9b.

Tabela 11 – Resultados parciais da identificação de especialistas em Ciência da Computação envolvendo o Score Propagated

(a) Propagated		(b) Alfa e Propagated	
Precision@n	Recall@n	Precision@n	Recall@n
P@1	0	P@1	1
P@5	0	P@5	0,8
P@10	0	P@10	0,5
P@25	0	P@25	0,44
P@50	0,02	P@50	0,44
P@100	0,12	P@100	0,35
P@250	0,2	P@250	0,276
P@500	0,214	P@500	0,198
R-precision:	0,10853	R-precision:	0,32558
No. Experts:	129	No. Experts:	129

(c) Beta e Propagated

Precision@n	Recall@n
P@1	0
P@5	0,6
P@10	0,8
P@25	0,56
P@50	0,54
P@100	0,48
P@250	0,332
P@500	0,208
R-precision:	0,44186
No. Experts:	129

Fonte: Jamison José da Silva Lima, 2017

A seguir será realizada a otimização dos parâmetros utilizados no Score Propagated, então serão realizados os testes para a seleção final dos parâmetros antes da apresentação dos resultados finais.

Tipo de média

A otimização dos parâmetros do Score Propagated foi iniciada pelo parâmetro binário *somaAlfaBeta*, que seleciona o tipo de média utilizada entre geométrica e aritmética para os valores utilizados como entrada do Score Propagated. Como os valores Alfa e Beta são previamente normalizados no intervalo $[0,1]$, para o cálculo da média geométrica é somado 1 a cada um dos Scores, e após a execução é subtraído 1 do resultado, para evitar que alguns autores com o Score Alfa ou Beta igual a 0 deixem de ser contabilizados.

A tabela 12 apresenta a variação na precisão dos resultados pela mudança desse parâmetro.

Tabela 12 – Variação da precisão dos resultados de acordo com o tipo de média utilizada

Área	Computação (R=129)		Química (R=286)		Física (R=435)	
Tipo Média	Aritmética	Geométrica	Aritmética	Geométrica	Aritmética	Geométrica
R-Precision	0,48062	0,52713	0,47552	0,49301	0,51034	0,51034
P@1	1	1	1	1	1	1
P@5	0,8	0,8	0,8	0,8	1	1
P@10	0,8	0,8	0,8	0,8	0,6	0,6
P@25	0,76	0,8	0,6	0,64	0,6	0,6
P@50	0,6	0,64	0,54	0,6	0,5	0,5
P@100	0,51	0,54	0,55	0,58	0,51	0,52
P@250	0,36	0,36	0,484	0,532	0,56	0,56
P@500	0,226	0,228	0,374	0,38	0,498	0,494

Fonte: Jamison José da Silva Lima, 2017

Analisando os resultados obtidos, é possível afirmar que a média geométrica aprimora os resultados obtidos em praticamente todos os pontos nos quais foi calculada a precisão, por isso essa média continuará sendo utilizada nos próximos testes.

Poda do grafo

Os próximos parâmetros analisados realizam a remoção de arestas envolvendo nós que não participam do conjunto que está sendo buscado e de nós da rede que não possuem relacionamentos. Se o parâmetro *subrede* estiver ativo, as arestas envolvendo autores que não declaram publicar na área são removidas da rede. Caso o parâmetro *removeOrfao* for ativado, os autores que não possuem nenhuma coautoria na rede são removidos.

Como esses parâmetros modificam a estrutura da rede, foram analisados os efeitos colaterais da utilização deles, para depois serem realizados os testes envolvendo a precisão dos resultados.

A partir de testes realizados envolvendo os especialistas das três áreas, foi identificado que a remoção de nós órfãos da rede apresenta problemas por ter removido três bolsistas nível 1 da Física, o que nos aponta que existem alguns pesquisadores com alto nível de especialidade que não possuem relações com outros pesquisadores de alto nível que também recebem essa bolsa, portanto esse parâmetro não será utilizado nos próximos testes. Já a poda da árvore por meio da remoção das relações entre pares de autores que não publicam na área não apresentou nenhum efeito problemático aos resultados, apenas reduzindo consideravelmente o tempo de execução do Score Propagated devido ao menor número de arestas na rede.

Após essa análise preliminar, são apresentados os resultados de acordo com a variação do parâmetro *subrede* na tabela 13.

Tabela 13 – Variação na precisão dos resultados a partir do parâmetro *subrede*

Área Subrede	Computação (R=129)		Química (R=286)		Física (R=435)	
	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
R-Precision	0,5271	0,5271	0,4930	0,4650	0,5103	0,5103
P@1	1	1	1	1	1	1
P@5	0,8	0,8	0,8	0,8	1	1
P@10	0,8	0,8	0,8	0,8	0,6	0,6
P@25	0,8	0,8	0,64	0,68	0,6	0,6
P@50	0,64	0,66	0,6	0,64	0,5	0,5
P@100	0,54	0,52	0,58	0,56	0,52	0,54
P@250	0,36	0,368	0,532	0,484	0,56	0,556
P@500	0,228	0,228	0,38	0,37	0,494	0,488
AP	0,5143	0,5134	0,4575	0,4402	0,5036	0,4967

Fonte: Jamison José da Silva Lima, 2017

É possível observar a partir dos resultados obtidos que apenas a utilização da métrica precisão não permite chegar a conclusões para a melhor variação desse parâmetro, com diversos pontos melhores para uma ou para outra opção do parâmetro. Porém, considerando a métrica *Average Precision*, foi identificado que a poda das arestas de outras áreas melhora um pouco a qualidade dos resultados, portanto será utilizada nos próximos testes.

Taxa de propagação e fator de parada

A taxa de propagação e o fator de parada são os últimos parâmetros que passaram pela otimização. A taxa de propagação d representa a taxa de transferência de especialidade entre dois vizinhos no intervalo $(0,1)$, de modo que quanto maior esse valor, maior a transmissão do Score Propagated entre os autores. Já o fator de parada *epsilon* limita o número de execuções do algoritmo de propagação de acordo com o erro médio da rede. Após a realização de testes preliminares para a identificação dos intervalos a serem analisados em detalhes, apresentaremos a seguir os resultados para o d no intervalo 0,01 a 0,20 e o *epsilon* entre 0,0001 e 0,01.

A figura 9 apresenta mapas de calor com os resultados para a área da Ciência da Computação.

A partir da análise desses gráficos, observa-se que os valores $d = 0,07$ e *epsilon* = $[0,003;0,01]$ maximizam a média entre as métricas AP@25 e *R-Precision*. O menor

Figura 9 – Resultados a partir da variação dos parâmetros d e ϵ para a Computação

Fonte: Jamison José da Silva Lima, 2017

valor de ϵ entre essas opções foi selecionado, isto é 0,003, minimizando a diferença entre o erro obtido com o ϵ . Aumentar este valor que está relacionado ao fator de parada implicaria numa maior permissividade do algoritmo ser executado em apenas uma iteração.

A figura 9c apresenta os resultados para a métrica *Average Precision*. Esta métrica é geralmente utilizada para medir a qualidade geral de algoritmos de busca. Comparando seus melhores resultados ($d = 0,02$ e $\epsilon = 0,0007$) aos obtidos anteriormente pela média entre as métricas *R-Precision* e AP@25, há um aumento de 0,03% no MAP e decréscimo de 5,21% no AP@25, enquanto o R-Precision mantém-se igual.

A *Average Precision* será utilizada para a seleção desses parâmetros para as áreas de Química e Física, conforme será apresentado a seguir.

A figura 10 apresenta os resultados para a área de Química.

Figura 10 – Resultados a partir da variação dos parâmetros d e ϵ para a Química

Fonte: Jamison José da Silva Lima, 2017

Em comparação aos resultados obtidos para a Computação, ocorre uma situação mais complexa para área da Química. Os melhores resultados do AP@25 divergem dos melhores valores do R -Precision e utilizar a média entre essas métricas não produz um resultado satisfatório, com $d = 0,02$ e o ϵ igual a 0,002.

A figura 10c apresenta os resultados da métrica *Average Precision* para a área de Química. A partir desses resultados, o melhor ponto encontrado envolve o $d = 0,01$ e o ϵ igual a 0,0004.

Comparando os resultados obtidos com o resultado da média da AP@25 com a R -Precision, houve um aumento de 3,95% na *Average Precision*, 6,16% na R -Precision e

uma redução de 59,8% na AP@25. Essa alta variação no AP@25 é um sinal de que essa não é uma métrica recomendável para otimizar os resultados dessa área.

As figuras 11b e 11a apresentam os resultados das métricas *R-Precision* e AP@25 para a área de Física.

Figura 11 – Resultados a partir da variação dos parâmetros d e ϵ para a Física

Fonte: Jamison José da Silva Lima, 2017

Para os resultados da Física ocorre o mesmo problema de divergência entre as métricas *R-Precision* e AP@25. O melhor valor para a média dessas duas métricas ocorre para o $d = 0,19$ e $\epsilon = 0,0001$.

A figura 11c apresenta o mapa de calor dos resultados da métrica *Average Precision* que possui como máximos o $d = 0,01$ e $\epsilon = 0,0002$. Em relação aos resultados envolvendo a média, há o aumento de 4,81% no AP e de 5,95% na *R-Precision*, com a redução de 79,7% no AP@25.

Apesar de consideravelmente melhor que a AP@10 para a análise do tamanho da ontologia como apresentado anteriormente, a métrica AP@25 não é estável o suficiente para a seleção desses parâmetros. Uma hipótese para esse problema envolve a maior quantidade de especialistas contidos nessas áreas (286 para a Química e 435 para a Física) em relação à Computação (com 129 especialistas).

Na tabela 14 são apresentados os três conjuntos de parâmetros selecionados para os testes finais.

Tabela 14 – Valores dos parâmetros utilizados após o treinamento

Score	Nome da variável	Descrição	Computação	Química	Física
Alfa	<i>areaMaxSize</i>	Número máximo de áreas para adicionar na ontologia	<i>1</i>		
	<i>ontologiaSize</i>	Tamanho da ontologia	<i>300</i>	<i>184</i>	<i>141</i>
Beta	<i>maxArea</i>	Normalizar a pontuação Qualis levando em consideração o valor máximo da área	<i>true</i>	<i>true</i>	<i>false</i>
	<i>maxAutor</i>	Normalizar a pontuação Qualis levando em consideração o valor máximo do autor	<i>true</i>		
Propagated	<i>somaAlfaBeta</i>	Se verdadeiro, calcula a média aritmética de Alfa e Beta, senão é calculada a média geométrica	<i>false</i>		
	<i>subrede</i>	Remove arestas entre autores que não publicaram na área	<i>true</i>		
	<i>removeOrfao</i>	Remove nós de autores que não possuem nenhuma aresta	<i>false</i>		
	<i>d</i>	Taxa de propagação no intervalo (0,1)	<i>0,07</i>	<i>0,01</i>	<i>0,01</i>
	<i>epsilon</i>	Erro médio utilizado como fator de parada	<i>0,003</i>	<i>0,0004</i>	<i>0,0002</i>

Fonte: Jamison José da Silva Lima, 2017

Após o final do treinamento, é possível reparar que diversos parâmetros se mostraram estáveis entre as áreas estudadas, restando apenas quatro com diferentes valores que serão levados em conta nos testes: o tamanho da ontologia, a utilização ou não da normalização por áreas, a taxa de propagação e o fator de parada. Estes resultados serão apresentados a seguir.

5.2.4 Comparação dos conjuntos de parâmetros

Após o treinamento realizado nas seções anteriores, foram selecionadas as melhores combinações de parâmetros a partir dos resultados para a métrica MAP, que foi escolhida por refletir a qualidade dos resultados para todos os especialistas, em especial os primeiros (NIST, 2013; SU; YUAN; ZHU, 2015).

Para os testes foram selecionadas 17 subáreas além das três utilizadas no treinamento: Matemática, Letras, Odontologia, Educação, História, Astronomia, Administração, Engenharia Civil, Geografia, Nutrição, Medicina, Psicologia, Zoologia, Zootecnia, Ecologia, Economia e Genética, representando 23% do número total de 88 áreas declaradas para esse conjunto.

Especialmente para a área de Medicina, foi feita uma pequena modificação na seleção da pontuação Qualis, pois esta área é subdividida em outras e neste trabalho foram utilizados apenas os estratos Qualis da Medicina III (que envolve Ginecologia, Oftalmologia, Ortopedia e Cirurgia).

A tabela 15 apresenta as 9 variações possíveis entre os parâmetros *tamListaPalavras* e para os pares *d* e *epsilon*, de modo que as três primeiras representam os parâmetros *tamListaPalavras* e *epsilon* selecionados anteriormente para as áreas de Ciência da Computação, Química e Física, respectivamente. Adicionalmente, para cada uma dessas combinações, o parâmetro *maxArea* pode receber os valores *true* ou *false*, totalizando 18 variações entre os parâmetros.

Tabela 15 – Lista de variações dos parâmetros para a realização dos próximos testes

Parâmetro	Conjunto de Parâmetros								
	1	2	3	4	5	6	7	8	9
<i>ontologySize</i>	300	184	141	300	184	141	300	184	141
<i>{d*100, epsilon*10^4}</i>	{7, 30}	{1, 4}	{1, 2}	{1, 4}	{1, 2}	{7, 30}	{1, 2}	{7, 30}	{1, 4}

Fonte: Jamison José da Silva Lima, 2017

A tabela 16 apresenta a Average Precision para cada uma das áreas e o Mean Average Precision para os conjuntos de parâmetros com *maxArea* verdadeiro. A tabela 17 apresenta os mesmos resultados variando o parâmetro *maxArea* para falso. O valor máximo de cada linha está destacado em negrito.

Nessas tabelas é possível reparar que para as áreas de Ciência da Computação e Química a otimização apresentou os melhores resultados para seus conjuntos de parâmetros originais, como pode ser visto a partir dos resultados mostrados na tabela 16. Entretanto, para os resultados da Física (conjunto de parâmetros 3 da tabela 17) isso não aconteceu, provavelmente pela variação do parâmetro de normalização por área, visto que o melhor resultado da área ocorre para o conjunto de parâmetros 5, que possui como diferença uma lista de palavras ligeiramente maior.

É possível notar que houve uma queda de quase 2 pontos percentuais nos resultados pela utilização da normalização a partir do máximo da área, e que entre os resultados

Tabela 16 – Average Precision para cada conjunto de parâmetros com a normalização por área ativada

Área	Conjunto de Parâmetros								
	1	2	3	4	5	6	7	8	9
Ciência da Computação	0,5197	0,4832	0,5103	0,4861	0,5140	0,5067	0,4916	0,5185	0,4636
Química	0,4431	0,4613	0,4536	0,4521	0,4494	0,4599	0,4290	0,4583	0,4593
Física	0,4674	0,4228	0,5163	0,4253	0,5208	0,4902	0,5114	0,4619	0,4612
Matemática	0,5267	0,4713	0,4647	0,4792	0,4940	0,4772	0,5471	0,5032	0,4377
Letras	0,1620	0,1715	0,1776	0,1628	0,1715	0,1819	0,1628	0,1723	0,1776
Odontologia	0,7376	0,6993	0,7249	0,7437	0,7081	0,7139	0,7196	0,7093	0,6852
Educação	0,3937	0,3808	0,3912	0,3908	0,3878	0,3883	0,3881	0,3863	0,3792
História	0,1843	0,1397	0,1226	0,1653	0,1397	0,1421	0,1653	0,1569	0,1226
Astronomia	0,2948	0,2954	0,3067	0,2818	0,2954	0,3360	0,2818	0,3222	0,3067
Administração	0,1919	0,1985	0,1997	0,2025	0,1985	0,1920	0,2025	0,1917	0,1997
Engenharia Civil	0,3907	0,3407	0,3769	0,3384	0,3981	0,3773	0,4026	0,3878	0,3360
Geografia	0,1768	0,1918	0,2393	0,1699	0,1918	0,2400	0,1699	0,1940	0,2393
Nutrição	0,3060	0,3172	0,3275	0,3266	0,3201	0,3208	0,3288	0,3085	0,3144
Medicina	0,3584	0,3782	0,3798	0,3652	0,3691	0,3821	0,3537	0,3744	0,3871
Psicologia	0,2697	0,2088	0,2260	0,2386	0,2330	0,2366	0,2735	0,2416	0,2023
Zoologia	0,2957	0,2651	0,2750	0,2866	0,2715	0,2749	0,2943	0,2738	0,2703
Zootecnia	0,5892	0,5762	0,5622	0,5727	0,5816	0,5783	0,5867	0,5945	0,5617
Ecologia	0,2783	0,3249	0,2968	0,3286	0,2822	0,3146	0,2824	0,3006	0,3490
Economia	0,2981	0,2807	0,2872	0,3045	0,2807	0,2958	0,3045	0,2786	0,2872
Genética	0,2956	0,3203	0,3186	0,3017	0,3073	0,3209	0,2942	0,3096	0,3334
Mean Average Precision	0,3590	0,3464	0,3578	0,3511	0,3557	0,3615	0,3595	0,3572	0,3487

Fonte: Jamison José da Silva Lima, 2017

Tabela 17 – Average Precision para cada conjunto de parâmetros com a normalização por área desativada

Área	Conjunto de Parâmetros								
	1	2	3	4	5	6	7	8	9
Ciência da Computação	0,4602	0,4096	0,4738	0,4464	0,4720	0,4663	0,4621	0,4669	0,3825
Química	0,3918	0,4111	0,4003	0,3992	0,4002	0,4003	0,3885	0,4036	0,4081
Física	0,5520	0,5565	0,5593	0,5507	0,5602	0,5568	0,5532	0,5589	0,5534
Matemática	0,5676	0,4868	0,5384	0,4858	0,5454	0,5283	0,5481	0,5484	0,4704
Letras	0,2527	0,2353	0,2476	0,2409	0,2353	0,2611	0,2409	0,2424	0,2476
Odontologia	0,5082	0,5476	0,5146	0,5895	0,5134	0,4968	0,5358	0,4940	0,5428
Educação	0,3925	0,3692	0,3935	0,3791	0,3898	0,3812	0,3920	0,3832	0,3607
História	0,2834	0,2232	0,2139	0,2484	0,2232	0,2534	0,2484	0,2678	0,2139
Astronomia	0,4585	0,4127	0,4309	0,4454	0,4127	0,4709	0,4454	0,4672	0,4309
Administração	0,2177	0,2017	0,2089	0,2117	0,2017	0,2162	0,2162	0,2093	0,2089
Engenharia Civil	0,4569	0,3771	0,4478	0,3726	0,4446	0,4483	0,4450	0,4521	0,3813
Geografia	0,3106	0,3554	0,3955	0,3114	0,3554	0,3610	0,3114	0,3338	0,3955
Nutrição	0,2741	0,3304	0,3150	0,3348	0,3051	0,2830	0,3117	0,2746	0,3450
Medicina	0,3302	0,3386	0,3390	0,3314	0,3368	0,3387	0,3299	0,3371	0,3388
Psicologia	0,3907	0,3071	0,3926	0,3157	0,3870	0,3839	0,3971	0,3790	0,3155
Zoologia	0,2579	0,2615	0,2514	0,2795	0,2498	0,2496	0,2599	0,2483	0,2637
Zootecnia	0,5404	0,5935	0,5754	0,5949	0,5876	0,5490	0,5838	0,5546	0,5793
Ecologia	0,2106	0,2424	0,2176	0,2428	0,2128	0,2147	0,2133	0,2105	0,2488
Economia	0,4078	0,3692	0,3606	0,3674	0,3692	0,4067	0,3674	0,4097	0,3606
Genética	0,2850	0,3002	0,2955	0,2927	0,2906	0,3002	0,2848	0,2904	0,3070
Mean Average Precision	0,3774	0,3665	0,3786	0,3720	0,3746	0,3783	0,3767	0,3766	0,3677

Fonte: Jamison José da Silva Lima, 2017

obtidos, pequenas variações nos conjuntos de parâmetros modificam significativamente os resultados, como a redução de mais de um ponto percentual (de 37,86% para 36,77%)

entre o melhor resultado e o conjunto 9 (no qual é realizada a variação do *epsilon* em 0,0002).

Além da variação da precisão média entre os diferentes conjuntos de parâmetros, existem grandes variações entre os melhores resultados de áreas diferentes, variando entre 0,2177 na área de Administração a 0,5949 para a Zootecnia, ambos resultados obtidos pela normalização por autor.

Com relação ao *Mean Average Precision*, os melhores resultados foram obtidos para o conjunto de parâmetros 3 (otimizados para a área da Física), com 141 palavras, *d* e *epsilon* iguais a 0,01 e 0,0002, e com a normalização por área desativada.

A tabela 18 apresenta os valores dos parâmetros que serão utilizados para a produção dos resultados finais.

Tabela 18 – Lista dos parâmetros selecionados após a otimização

Score	Nome da variável	Descrição	Valor
Alfa	<i>areaMaxSize</i>	Número máximo de áreas para adicionar na ontologia	<i>1</i>
	<i>ontologiaSize</i>	Tamanho mínimo da ontologia	<i>141</i>
Beta	<i>maxArea</i>	Normalizar a pontuação Qualis levando em consideração o valor máximo para a área	<i>false</i>
	<i>maxAutor</i>	Normalizar a pontuação Qualis levando em consideração o valor máximo para o autor	<i>true</i>
Propagated	<i>somaAlfaBeta</i>	Se verdadeiro, calcula a média aritmética de Alfa e Beta, senão é calculada a média geométrica	<i>false</i>
	<i>subrede</i>	Remove arestas entre autores que não publicaram na área	<i>true</i>
	<i>removeOrfao</i>	Remove nós de autores que não possuem nenhuma aresta	<i>false</i>
	<i>d</i>	Taxa de propagação no intervalo (0,1)	<i>0,01</i>
	<i>epsilon</i>	Erro médio utilizado como fator de parada	<i>0,0002</i>

Fonte: Jamison José da Silva Lima, 2017

5.3 Resultados finais

Nessa seção serão apresentados os resultados finais de acordo com os parâmetros otimizados anteriormente, envolvendo as métricas *Precision*, *Recall*, *R-Precision* e *Mean*

Average Precision, levando em conta o modelo utilizado na *Text REtrieval Conference*¹ (NIST, 2013).

Os resultados serão comparados utilizando uma versão adaptada do algoritmo implementado por Chagas, Perez-Alcazar e Digiampietri (2015), que foi a base para esse projeto. As principais diferenças desse algoritmo são: os Scores Alfa e Beta são obtidos e normalizados para cada publicação, sendo repassados como valor inicial para o Score Propagated de cada autor. Não são utilizadas estratégias para a poda do grafo, sendo assim utilizado o conjunto completo de candidatos e coautorias.

5.3.1 Resultados do conjunto de Bolsistas Produtividade

A tabela 19 apresenta os resultados para as métricas Recall Level Precision, Mean Average Precision, Precision@n e R-Precision médios obtidos para o conjunto de Bolsistas Produtividade.

Tabela 19 – Resultados para o conjunto de Bolsistas Produtividade

(a) Recall Level Precision Averages			(b) Document Level Averages		
Recall Level	Precision		Document Level	Precision	
	Base	Proposto		Base	Proposto
0,00	0,3000	0,2500	At 5 docs	0,2200	0,3600
0,10	0,2277	0,4509	At 10 docs	0,2250	0,3800
0,20	0,2001	0,4479	At 15 docs	0,2267	0,3933
0,30	0,1726	0,4234	At 20 docs	0,2325	0,4075
0,40	0,1545	0,4176	At 25 docs	0,2280	0,4120
0,50	0,1362	0,3925	At 30 docs	0,2100	0,4150
0,60	0,1187	0,3810	At 50 docs	0,1950	0,4290
0,70	0,0993	0,3550	At 100 docs	0,1890	0,3950
0,80	0,0768	0,3260	At 200 docs	0,1613	0,3505
0,90	0,0565	0,2752	At 500 docs	0,1168	0,2185
1,00	0,0232	0,1283	At 1000 docs	0,0823	0,1262
MAP	0,1399	0,3786	R-Precision	0,1751	0,4047

Fonte: Jamison José da Silva Lima, 2017

A partir desses resultados é possível reparar que há um aprimoramento significativo da abordagem proposta em relação ao algoritmo utilizado como base, sendo um bom sinal de que a abordagem proposta aprimora a identificação de especialistas em conjuntos de dados contendo candidatos de diversas áreas.

¹ <http://trec.nist.gov>

5.3.2 Resultados do conjunto de Inteligência Artificial

A tabela 20 apresenta os resultados para as métricas *Recall Level Precision Averages*, *Average Precision*, *Precisão* e *R-Precision*. Nos testes realizados foram usados os mesmos parâmetros otimizados para o conjunto de Bolsistas Produtividade, e os resultados foram comparados com a lista de especialistas obtida a partir da realização do questionário que foi descrito anteriormente na seção 1.2.

Tabela 20 – Resultados para o conjunto de Inteligência Artificial

(a) Recall Level Precision

Recall Level	Precision		
	Base	Proposto	Computação
0,00	1,0000	1,0000	1,0000
0,10	0,7500	1,0000	1,0000
0,20	0,8333	1,0000	0,7143
0,30	0,6154	0,3810	0,5000
0,40	0,6250	0,4348	0,4167
0,50	0,5000	0,3429	0,2264
0,60	0,3488	0,3000	0,2239
0,70	0,3617	0,3148	0,1667
0,80	0,1639	0,1429	0,0557
0,90	0,0940	0,0440	0,0241
1,00	0,0221	0,0173	0,0141
AP	0,4934	0,4463	0,3821

(b) Document Level Precision

Document Level	Precision		
	Base	Proposto	Computação
At 5 docs	0,8000	1,0000	0,8000
At 10 docs	0,7000	0,6000	0,5000
At 15 docs	0,6000	0,4000	0,4667
At 20 docs	0,5500	0,3500	0,4500
At 25 docs	0,4800	0,4000	0,4000
At 30 docs	0,4333	0,3667	0,3667
At 50 docs	0,3600	0,3000	0,2200
At 100 docs	0,1900	0,1900	0,1600
At 200 docs	0,1050	0,1000	0,0950
At 500 docs	0,0440	0,0440	0,0400
At 1000 docs	0,0230	0,0230	0,0220
R-Precision	0,5000	0,4167	0,4167

Fonte: Jamison José da Silva Lima, 2017

Diferentemente dos testes realizados para o conjunto de Bolsistas Produtividade, os resultados foram melhores para o algoritmo base em diversos pontos, principalmente para o *Recall* 0,3 e 0,5, que representa o intervalo entre 7 a 12 pesquisadores encontrados. O único ponto de destaque para o algoritmo proposto ocorre para os 5 primeiros resultados, no qual a precisão é máxima (100%), contra 80% do sistema utilizado como base.

Também foram apresentados os resultados envolvendo o conjunto de parâmetros otimizados para a área de Ciência da Computação pela proximidade com a Inteligência Artificial. Os parâmetros selecionados pelos testes superam os valores obtidos para os bolsistas produtividade da Ciência da Computação.

5.3.3 Resultados do conjunto de Nanotecnologia

Para o conjunto de Nanotecnologia foram reproduzidos os testes realizados por Chagas, Perez-Alcazar e Digiampietri (2015), que consistiram na comparação entre pares de autores por meio de questionários com objetivo de indicar o principal especialista

em um assunto da área. Como neste projeto foram coletados apenas os resultados do questionário realizado, alguns detalhes da implementação podem estar diferentes, pois a lista de palavras obtida para a área foi adaptada como uma lista de palavras, para se tornar utilizável pelo algoritmo proposto.

O principal teste realizado em seu trabalho envolveu a posição relativa entre os pares de pesquisadores analisados manualmente. Para cada um dos 63 testes realizados (3 comparações foram invalidadas pois uma delas era repetida e outras duas eram conflitantes entre si), caso a ordem do par de pesquisadores obtida pelo algoritmo seja a mesma do teste realizado, é computado um acerto.

Os resultados obtidos nesse teste são apresentados na tabela 21.

Tabela 21 – Taxa de acertos da posição relativa para o conjunto de Nanotecnologia

	Acertos	Erros
Base	45 (71,43%)	18 (28,57%)
Proposto	46 (73,02%)	17 (26,98%)

Fonte: Jamison José da Silva Lima, 2017

A partir desses resultados é possível reparar que há uma pequena vantagem para os resultados obtidos pela abordagem proposta. Outro detalhe relevante é que o algoritmo base obteve resultados ligeiramente melhores aos obtidos originalmente por Chagas, Perez-Alcazar e Digiampietri (2015). Essa diferença deve ter ocorrido devido a atualização dos dados ou as diferenças em alguns detalhes da implementação do algoritmo, que levou em conta apenas a utilização de unigramas e que não possuía os mesmos valores para os parâmetros utilizados nesses testes.

Além do teste envolvendo a acurácia do algoritmo, também foram realizados testes levando em consideração o conjunto completo de pesquisadores citados no questionário realizado, de modo que todos os 27 pesquisadores que foram considerados relevantes nos questionários respondidos formaram a lista de especialistas utilizada nos próximos testes.

A tabela 22 apresenta os resultados para as métricas *Recall Level Precision*, *Average Precision*, *Precision* e *R-Precision* envolvendo a lista de especialistas gerada para a área e utilizando uma versão adaptada da lista de termos proposta por Aluísio et al. (2005), de modo que a ontologia foi traduzida para o português e adaptada para a estrutura de unigramas utilizada nesse trabalho.

Tabela 22 – Resultados para o conjunto de Nanotecnologia com a lista de palavras gerada *a priori*

(a) Recall Level Precision

Recall Level	Precision	
	Base	Proposto
0,00	1,0000	1,0000
0,10	1,0000	1,0000
0,20	1,0000	0,4615
0,30	0,3214	0,3333
0,40	0,1833	0,2391
0,50	0,2000	0,2593
0,60	0,1700	0,1683
0,70	0,1776	0,0742
0,80	0,1063	0,0566
0,90	0,0865	0,0478
1,00	0,0618	0,0367
AP	0,3892	0,3077

(b) Document Level Precision

Document Level	Precision	
	Base	Proposto
At 5 docs	1,0000	0,8000
At 10 docs	0,7000	0,4000
At 15 docs	0,5333	0,4000
At 20 docs	0,4000	0,3000
At 25 docs	0,3200	0,3200
At 30 docs	0,3000	0,3000
At 50 docs	0,2000	0,2400
At 100 docs	0,1700	0,1600
At 200 docs	0,1050	0,0900
At 500 docs	0,0540	0,0460
At 750 docs	0,0360	0,0360
R-Precision	0,2963	0,3333

Fonte: Jamison José da Silva Lima, 2017

A tabela 23 apresenta os resultados para as mesmas métricas, com a diferença da lista de palavras ter sido obtida automaticamente a partir da técnica proposta neste projeto.

Tabela 23 – Resultados para o conjunto de Nanotecnologia com a lista de palavras obtida automaticamente

(a) Recall Level Precision

Recall Level	Precision	
	Base	Proposto
0,00	1,0000	1,0000
0,10	1,0000	1,0000
0,20	1,0000	0,4615
0,30	0,3600	0,2727
0,40	0,2000	0,2619
0,50	0,2059	0,2222
0,60	0,1735	0,2394
0,70	0,1792	0,1338
0,80	0,1106	0,0561
0,90	0,0877	0,0472
1,00	0,0619	0,0385
AP	0,3917	0,3168

(b) Document Level Precision

Document Level	Precision	
	Base	Proposto
At 5 docs	1,0000	0,8000
At 10 docs	0,7000	0,4000
At 15 docs	0,5333	0,4667
At 20 docs	0,4000	0,3500
At 25 docs	0,3600	0,3200
At 30 docs	0,3000	0,2667
At 50 docs	0,1800	0,2400
At 100 docs	0,1700	0,1700
At 200 docs	0,1100	0,1000
At 500 docs	0,0540	0,0480
At 750 docs	0,0360	0,0360
R-Precision	0,3333	0,2963

Fonte: Jamison José da Silva Lima, 2017

Assim como ocorrido no conjunto de Inteligência Artificial, é possível notar uma vantagem na precisão dos resultados para o algoritmo base na maior parte dos pontos calculados.

Além disso, observa-se que a lista de palavras gerada a partir dos termos mais utilizados apresentou resultados ligeiramente melhores aos obtidos pelo uso da lista de palavras proveniente da ontologia proposta por Aluísio et al. (2005) em diversos pontos, sendo esse um indício de que a geração automática das listas de palavras possui potencial para bons resultados, além de não necessitar de uma ontologia de domínio de referência gerada *a priori*.

O último teste realizado envolveu a comparação entre a lista de palavras obtidas a partir da abordagem proposta nesse projeto com a ontologia gerada por Aluísio et al. (2005). Após comparar as 144 palavras obtidas pela abordagem proposta, 77,08% delas pertenciam a ontologia gerada por Aluísio et al. (2005).

5.4 Considerações

Este capítulo apresentou a otimização dos parâmetros e os resultados obtidos pela abordagem desenvolvida, os quais foram comparados com os resultados do algoritmo base para esse projeto, proposto por Chagas, Perez-Alcazar e Digiampietri (2015).

A tabela 24 apresenta o Average Precision para cada composição dos Scores propostos.

O algoritmo proposto possui a melhor precisão média geral (MAP), além de apresentar os melhores resultados (em negrito) para 7 das 20 áreas testadas (resultado superior a qualquer uma das medidas individualmente ou as demais combinações das medidas), indicando que a composição proposta aprimora os resultados. Com 4,2 pontos percentuais a menos no MAP, a composição *alfa+beta* obteve o segundo lugar nos resultados médios. Já a composição *alfa+propagated* obteve o melhor resultado da métrica AP em 5 das 20 áreas testadas. Por outro lado, os piores resultados médios ficaram com a execução solo do Score Propagated.

Os resultados para o conjunto de bolsistas produtividade apresentam uma boa vantagem em relação aos resultados do algoritmo base, sendo um bom indicativo de maior precisão para esse tipo de conjunto. Por outro lado, os resultados dos outros conjuntos de dados foram ligeiramente melhores para o algoritmo base. É possível que a obtenção dos Score Alfa e Beta para publicações apresente melhores resultados caso todo o conjunto de pesquisadores seja da mesma área. Mas é importante ressaltar que esses conjuntos

Tabela 24 – Average Precision para cada composição do conjunto de Scores

Área	Composição do algoritmo						
	alfa	beta	alfa+beta	prop	alfa+prop	beta+prop	full
Ciência da Computação	0,2320	0,3477	0,3109	0,1616	0,3300	0,4607	0,4738
Química	0,3096	0,2052	0,3227	0,2497	0,4105	0,3461	0,4003
Física	0,3607	0,3525	0,4395	0,3111	0,4985	0,5001	0,5593
Matemática	0,4470	0,2845	0,4643	0,1035	0,4677	0,3983	0,5384
Letras	0,0971	0,1356	0,2489	0,0572	0,0980	0,1792	0,2476
Odontologia	0,4182	0,3344	0,4395	0,3650	0,5329	0,4659	0,5146
Educação	0,3012	0,2953	0,3573	0,0580	0,3272	0,3252	0,3935
História	0,0801	0,1851	0,2160	0,0769	0,0811	0,2436	0,2139
Astronomia	0,4170	0,0749	0,3959	0,2141	0,4247	0,1100	0,4309
Administração	0,1009	0,1527	0,2068	0,0949	0,1015	0,2171	0,2089
Engenharia Civil	0,2977	0,2704	0,3708	0,2769	0,3321	0,4109	0,4478
Geografia	0,2086	0,1317	0,4041	0,0891	0,2066	0,2184	0,3955
Nutrição	0,2598	0,1352	0,2921	0,1810	0,3273	0,2214	0,3150
Medicina	0,2140	0,1711	0,2261	0,2086	0,3473	0,2986	0,3390
Psicologia	0,0754	0,3018	0,3203	0,1925	0,0968	0,3986	0,3926
Zoologia	0,2372	0,1806	0,2674	0,1712	0,2500	0,2024	0,2514
Zootecnia	0,4579	0,1900	0,5304	0,2792	0,4863	0,2740	0,5754
Ecologia	0,2546	0,0780	0,3102	0,1295	0,2853	0,1494	0,2176
Economia	0,1753	0,3219	0,3568	0,1186	0,1776	0,4702	0,3606
Genética	0,2553	0,0870	0,2427	0,1955	0,3172	0,2531	0,2955
Mean Average Precision	0,2600	0,2118	0,3361	0,1767	0,3049	0,3072	0,3786

Fonte: Jamison José da Silva Lima, 2017

possuem amostras pequenas em relação a população, portanto esses resultados não são conclusivos. Se houverem conjuntos de dados maiores disponíveis para treinamento dos parâmetros ou maiores conjuntos de especialistas, a solução proposta teria condições de superar os resultados da solução base para uma grande gama de domínios diferentes.

Também foi testada a relações entre serem considerados especialistas de Inteligência Artificial (por meio dos questionários) e o recebimento de Bolsa Produtividade. Observou-se que 8 dos pesquisadores citados recebiam Bolsa Produtividade nível 1 e 8 deles possuem Bolsa Produtividade nível 2 em Ciência da Computação. Além disso, outros 3 pesquisadores recebem essa bolsa para a área de Engenharia Elétrica, sendo duas delas nível 1. Esses resultados são um indício de que a Bolsa Produtividade CNPq possui uma alta correlação com a ideia de especialidade.

Todos os dados utilizados nesse projeto estão disponíveis na Internet².

² goo.gl/FiFTEo

6 Conclusões

Esse trabalho teve como objetivo a identificação de especialistas nas redes sociais acadêmicas a partir da composição envolvendo técnicas de recuperação de informação com um algoritmo de propagação. Para a realização dos testes e validação foram utilizados dados extraídos da Plataforma Lattes.

Foi proposto, desenvolvido e testado um algoritmo que levou em consideração o conteúdo textual dos títulos das publicações dos pesquisadores e os estratos Qualis dos veículos nos quais os artigos foram publicados, além do número e relevância de seus coautores.

Foram coletados dados de 17.387 pesquisadores divididos em três conjuntos diferentes: o conjunto de Bolsistas Produtividade em Pesquisa do CNPq, pesquisadores que atuam na área de Inteligência Artificial e pesquisadores que atuam em Nanotecnologia, incluindo cerca de 1.2 milhão de títulos de publicações.

Realizou-se uma otimização dos 9 parâmetros propostos, buscando maximizar a precisão dos resultados. Finalmente, foram realizados testes comparando a abordagem proposta com o algoritmo criado por Chagas, Perez-Alcazar e Digiampietri (2015).

Para o conjunto de Bolsistas Produtividade, a validação foi feita comparando os resultados obtidos ao recebimento da bolsa nível 1. Para os conjuntos de Inteligência Artificial e Nanotecnologia a validação foi realizada comparando os resultados obtidos pelos algoritmos com os especialistas citados nos questionários por pesquisadores atuantes da área.

6.1 Contribuições

As principais contribuições deste projeto foram as seguintes:

- Os dados e resultados obtidos nesse trabalho poderão ser reaproveitados para a comparação com outras técnicas de identificação de especialistas.
- Por meio dos testes realizados envolvendo os resultados parciais da identificação de especialistas para o conjunto de Bolsistas Produtividade, foi possível verificar que a composição de abordagens aprimorou os resultados.

- Foram encontrados indícios de que o algoritmo proposto possui potencial de identificar especialistas principalmente em conjuntos envolvendo várias áreas, obtendo o resultado do *Mean Average Precision* de 0,3786 para os testes envolvendo o conjunto de Bolsistas Produtividade, em comparação aos 0,1399 do algoritmo base.
- A abordagem automática para a geração da lista de palavras proposta nesse projeto, a qual é utilizada para inferir a importância dos títulos, obteve resultados parecidos aos obtidos pela ontologia proposta por Aluísio et al. (2005) para a área de Nanotecnologia, de modo que 77,08% das palavras obtidas pertenciam a ontologia gerada.

6.2 Trabalhos futuros

Existem diversos pontos que poderão ser aprimorados em trabalhos futuros a partir das observações realizadas e dos resultados obtidos nesse projeto:

- Comparar diretamente diferentes abordagens para a identificação de especialistas.
- Aprofundar a comparação de técnicas focadas em dados da Plataforma Lattes.
- Utilizar técnicas de recuperação de informação mais complexas, como os modelos estatísticos ESA e LDA.
- Aumentar as informações utilizadas para a identificação dos especialistas, como resumos e artigos completos, formação acadêmica e locais de trabalho.
- Incluir novas etapas na composição das abordagens, como a regressão dos resultados parciais.
- Aprimorar a validação dos resultados e a análise da correlação entre especialidade e recebimento de Bolsa Produtividade a partir da aplicação de questionários com mais pesquisadores.

Referências¹

- ALUÍSIO, S. M. et al. *Desenvolvimento de uma estrutura conceitual (ontologia) para a área de Nanociência e Nanotecnologia*. [S.l.], 2005. Disponível em: http://www.inf.pucrs.br/~ontolp/downloads/Ontonano/RT_276.pdf. Citado 4 vezes nas páginas 18, 72, 74 e 77.
- BARBOSA, C.; SOUZA, J. Moreira de; OLIVEIRA, J. Empowering the Delphi decision-making process using expert search from social networks. In: *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*. [S.l.: s.n.], 2014. p. 630–635. Citado 2 vezes nas páginas 16 e 39.
- BOK, K. et al. An expert search scheme using user activities and reliabilities in social networks. In: *2015 International Conference on Big Data and Smart Computing (BigComp)*. [S.l.: s.n.], 2015. p. 157–161. Citado 3 vezes nas páginas 16, 31 e 38.
- BONCHI, F. et al. Social Network Analysis and Mining for Business Applications. *ACM Trans. Intell. Syst. Technol.*, v. 2, n. 3, p. 22:1–22:37, maio 2011. ISSN 2157-6904. Disponível em: <http://doi.acm.org/10.1145/1961189.1961194>. Citado na página 38.
- BOSCH, A. v. d.; BOGERS, T.; KUNDER, M. d. Estimating search engine index size variability: a 9-year longitudinal study. *Scientometrics*, v. 107, n. 2, p. 839–856, maio 2016. ISSN 0138-9130, 1588-2861. Disponível em: <https://link.springer.com/article/10.1007/s11192-016-1863-z>. Citado na página 15.
- BOZZON, A. et al. Choosing the Right Crowd: Expert Finding in Social Networks. In: *Proceedings of the 16th International Conference on Extending Database Technology*. New York, NY, USA: ACM, 2013. (EDBT '13), p. 637–648. ISBN 978-1-4503-1597-5. Disponível em: <http://doi.acm.org/10.1145/2452376.2452451>. Citado na página 35.
- CAMPBELL, C. S. et al. Expertise Identification Using Email Communications. In: *Proceedings of the Twelfth International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2003. (CIKM '03), p. 528–531. ISBN 978-1-58113-723-1. Disponível em: <http://doi.acm.org/10.1145/956863.956965>. Citado 2 vezes nas páginas 26 e 32.
- CHAGAS, F. d. M.; PEREZ-ALCAZAR, J. D. J.; DIGIAMPIETRI, L. A. Algoritmo de classificação de especialistas em áreas na base de currículos Lattes. *Em Questão*, v. 21, n. 2, p. 119–139, set. 2015. ISSN 1808-5245. Disponível em: <http://www.seer.ufrgs.br/index.php/EmQuestao/article/view/52652>. Citado 17 vezes nas páginas 15, 16, 17, 18, 35, 40, 41, 43, 45, 47, 49, 50, 70, 71, 72, 74 e 76.
- CHEN, J.-M. et al. Efficient top-k support documents for expert search using relationship in a social network. In: *2011 International Conference on Machine Learning and Cybernetics (ICMLC)*. [S.l.: s.n.], 2011. v. 4, p. 1479–1484. Citado 2 vezes nas páginas 16 e 33.
- CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO. *RN-028/2015*. 2015. Disponível em: http://cnpq.br/view/-/journal_content/56_INSTANCE_0oED/10157/2958271?COMPANY_ID=10132. Citado na página 17.

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

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO. *Painel Lattes*. 2016. Disponível em: <http://estatico.cnpq.br/painelLattes/>. Citado na página 15.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR. *Classificação da produção intelectual*. 2016. Disponível em: <http://www.capes.gov.br/avaliacao/instrumentos-de-apoio/classificacao-da-producao-intelectual>. Citado na página 17.

DENG, H. et al. Modeling and Exploiting Heterogeneous Bibliographic Networks for Expertise Ranking. In: *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY, USA: ACM, 2012. (JCDL '12), p. 71–80. ISBN 978-1-4503-1154-0. Disponível em: <http://doi.acm.org/10.1145/2232817.2232833>. Citado 2 vezes nas páginas 34 e 36.

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 na página 15.

DIGIAMPIETRI, L. A. et al. Minerando e caracterizando dados de currículos lattes. In: *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. Curitiba, PR, Brazil: [s.n.], 2012. p. 7. Citado 2 vezes nas páginas 15 e 24.

DIGIAMPIETRI, L. A. et al. BraX-Ray: An X-Ray of the Brazilian Computer Science Graduate Programs. *PLOS ONE*, v. 9, n. 4, p. e94541, abr. 2014. ISSN 1932-6203. Disponível em: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0094541>. Citado 3 vezes nas páginas 18, 24 e 42.

DIGIAMPIETRI, L. A. et al. Extração, caracterização e análises de dados de currículos Lattes. *Revista Eletrônica de Sistemas de Informação ISSN 1677-3071 doi:10.21529/RESI*, v. 14, n. 2, ago. 2015. ISSN 1677-3071. Disponível em: <http://www.periodicosibepes.org.br/ojs/index.php/reinfo/article/view/1528>. Citado 2 vezes nas páginas 15 e 24.

FAZEL-ZARANDI, M. et al. Expert Recommendation Based on Social Drivers, Social Network Analysis, and Semantic Data Representation. In: *Proceedings of the 2Nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems*. New York, NY, USA: ACM, 2011. (HetRec '11), p. 41–48. ISBN 978-1-4503-1027-7. Disponível em: <http://doi.acm.org/10.1145/2039320.2039326>. Citado na página 39.

GOH, O. S.; FUNG, C. C. Automated Knowledge Extraction from Internet for a Crisis Communication Portal. In: WANG, L.; JIN, Y. (Ed.). *Fuzzy Systems and Knowledge Discovery*. Springer Berlin Heidelberg, 2005, (Lecture Notes in Computer Science, 3614). p. 1226–1235. ISBN 978-3-540-28331-7 978-3-540-31828-6. Disponível em: http://link.springer.com/chapter/10.1007/11540007_162. Citado na página 15.

HEIL, S.; WILD, S.; GAEDKE, M. Collaborative Adaptive Case Management with Linked Data. In: *Proceedings of the 23rd International Conference on World Wide Web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2014. (WWW '14 Companion), p. 99–102. ISBN 978-1-4503-2745-9. Disponível em: <http://dx.doi.org/10.1145/2567948.2577030>. Citado na página 39.

JOUNG, Y.-J. et al. A Comparative Study of Expert Search Strategies in Online Social Networks. In: *2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA)*. [S.l.: s.n.], 2013. p. 960–967. Citado 7 vezes nas páginas 15, 21, 26, 29, 30, 31 e 38.

KARDAN, A.; MOHTAJ, S. A model for expert finding based on social network structure and underlying information diffusion network. In: *2013 5th Conference on Information and Knowledge Technology (IKT)*. [S.l.: s.n.], 2013. p. 472–477. Citado 2 vezes nas páginas 16 e 31.

KARDAN, A.; OMIDVAR, A.; FARAHMANDNIA, F. Expert finding on social network with link analysis approach. In: *2011 19th Iranian Conference on Electrical Engineering (ICEE)*. [S.l.: s.n.], 2011. p. 1–6. Citado na página 30.

KHAN, A.; WU, Y.; YAN, X. Emerging Graph Queries in Linked Data. In: *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*. [S.l.: s.n.], 2012. p. 1218–1221. Citado na página 38.

KLEINBERG, J. M. Authoritative Sources in a Hyperlinked Environment. *J. ACM*, v. 46, n. 5, p. 604–632, set. 1999. ISSN 0004-5411. Disponível em: <http://doi.acm.org/10.1145/324133.324140>. Citado 4 vezes nas páginas 21, 22, 26 e 30.

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 na página 15.

LI, N. et al. Experts Ranking on the Enterprise Microblogging Based on the PageRank Algorithm. In: *Services Computing Conference (APSCC), 2012 IEEE Asia-Pacific*. [S.l.: s.n.], 2012. p. 345–349. Citado na página 32.

LI, X. et al. A Service Mode of Expert Finding in Social Network. In: *2013 International Conference on Service Sciences (ICSS)*. [S.l.: s.n.], 2013. p. 220–223. Citado 3 vezes nas páginas 21, 25 e 35.

LI, Y. et al. Expertise Network Discovery via Topic and Link Analysis in Online Communities. In: *2012 IEEE 12th International Conference on Advanced Learning Technologies (ICALT)*. [S.l.: s.n.], 2012. p. 311–315. Citado na página 33.

LIU, J.-H.; LI, Y.-M. Design of Contextual Local Expert Support Mechanism. In: *Proceedings of the 17th International Conference on Electronic Commerce 2015*. New York, NY, USA: ACM, 2015. (ICEC '15), p. 17:1–17:5. ISBN 978-1-4503-3461-7. Disponível em: <http://doi.acm.org/10.1145/2781562.2781581>. Citado na página 35.

LIU, D. et al. Influence Analysis Based Expert Finding Model and Its Applications in Enterprise Social Network. In: *Services Computing (SCC), 2013 IEEE International Conference on*. [S.l.: s.n.], 2013. p. 368–375. Citado na página 33.

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 2 vezes nas páginas 15 e 24.

MITRE. Expert finding systems. In: MAYBURY, M. T. (Ed.). [S.l.: s.n.], 2006. p. 1–35. Citado na página 15.

MOVSHOVITZ-ATTIAS, D. et al. Analysis of the Reputation System and User Contributions on a Question Answering Website: StackOverflow. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. New York, NY, USA: ACM, 2013. (ASONAM '13), p. 886–893. ISBN 978-1-4503-2240-9. Disponível em: <http://doi.acm.org/10.1145/2492517.2500242>. Citado 3 vezes nas páginas 26, 31 e 38.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. *Common Evaluation Measures*. 2013. Disponível em: <http://trec.nist.gov/pubs/trec22/appendices/measures.pdf>. Citado 2 vezes nas páginas 66 e 70.

NEWMAN, M. E. J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, v. 98, n. 2, p. 404–409, jan. 2001. ISSN 0027-8424, 1091-6490. Disponível em: <http://www.pnas.org/content/98/2/404>. Citado na página 21.

PADHYE, R.; MUKHERJEE, D.; SINHA, V. S. API As a Social Glue. In: *Companion Proceedings of the 36th International Conference on Software Engineering*. New York, NY, USA: ACM, 2014. (ICSE Companion 2014), p. 516–519. ISBN 978-1-4503-2768-8. Disponível em: <http://doi.acm.org/10.1145/2591062.2591115>. Citado na página 31.

PAGE, L. et al. Techreport, *The PageRank Citation Ranking: Bringing Order to the Web*. 1999. Disponível em: <http://ilpubs.stanford.edu:8090/422/>. Citado 4 vezes nas páginas 21, 22, 23 e 30.

PUTTAN, C.; SENIVONGSE, T. Expertise ranking in question-answer social network groups. In: *Computer Science and Software Engineering (JCSSE), 2013 10th International Joint Conference on*. [S.l.: s.n.], 2013. p. 177–182. Citado 2 vezes nas páginas 25 e 34.

RIZOVA, P. S. *Are You Networked for Successful Innovation?* Rochester, NY, 2008. Disponível em: <https://papers.ssrn.com/abstract=1093112>. Citado na página 15.

SHAHRIARI, M.; PAREKODI, S.; KLAMMA, R. Community-aware Ranking Algorithms for Expert Identification in Question-answer Forums. In: *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*. New York, NY, USA: ACM, 2015. (i-KNOW '15), p. 8:1–8:8. ISBN 978-1-4503-3721-2. Disponível em: <http://doi.acm.org/10.1145/2809563.2809592>. Citado na página 31.

SILVA, J. da et al. An Extensible Service for Experts Recommendation on Distributed Software Development Projects. In: *2012 IEEE Seventh International Conference on Global Software Engineering Workshops (ICGSEW)*. [S.l.: s.n.], 2012. p. 18–21. Citado 2 vezes nas páginas 16 e 39.

SMIRNOVA, E. A Model for Expert Finding in Social Networks. In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2011. (SIGIR '11), p. 1191–1192. ISBN 978-1-4503-0757-4. Disponível em: <http://doi.acm.org/10.1145/2009916.2010114>. Citado na página 34.

SOUZA, C. et al. Routing Questions in Twitter: An Effective Way to Qualify Peer Helpers. In: *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*. [S.l.: s.n.], 2013. v. 1, p. 109–114. Citado 2 vezes nas páginas 16 e 33.

SRBA, I.; GRZNAR, M.; BIELIKOVA, M. Utilizing Non-QA Data to Improve Questions Routing for Users with Low QA Activity in CQA. In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. New York, NY, USA: ACM, 2015. (ASONAM '15), p. 129–136. ISBN 978-1-4503-3854-7. Disponível em: <http://doi.acm.org/10.1145/2808797.2809331>. Citado 3 vezes nas páginas 25, 32 e 38.

SU, W.; YUAN, Y.; ZHU, M. A Relationship Between the Average Precision and the Area Under the ROC Curve. In: *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. New York, NY, USA: ACM, 2015. (ICTIR '15), p. 349–352. ISBN 978-1-4503-3833-2. Disponível em: <http://doi.acm.org/10.1145/2808194.2809481>. Citado 3 vezes nas páginas 36, 54 e 66.

WASSERMAN, S.; FAUST, K. *Social network analysis: Methods and applications*. [S.l.]: Cambridge university press, 1994. Citado na página 21.

WELSER, H. T. et al. Finding Social Roles in Wikipedia. In: *Proceedings of the 2011 iConference*. New York, NY, USA: ACM, 2011. (iConference '11), p. 122–129. ISBN 978-1-4503-0121-3. Disponível em: <http://doi.acm.org/10.1145/1940761.1940778>. Citado na página 34.

YENITERZI, R.; CALLAN, J. Constructing Effective and Efficient Topic-specific Authority Networks for Expert Finding in Social Media. In: *Proceedings of the First International Workshop on Social Media Retrieval and Analysis*. New York, NY, USA: ACM, 2014. (SoMeRA '14), p. 45–50. ISBN 978-1-4503-3022-0. Disponível em: <http://doi.acm.org/10.1145/2632188.2632208>. Citado na página 34.

ZHANG, J.; ACKERMAN, M. S.; ADAMIC, L. Expertise Networks in Online Communities: Structure and Algorithms. In: *Proceedings of the 16th International Conference on World Wide Web*. New York, NY, USA: ACM, 2007. (WWW '07), p. 221–230. ISBN 978-1-59593-654-7. Disponível em: <http://doi.acm.org/10.1145/1242572.1242603>. Citado 2 vezes nas páginas 30 e 34.

ZHANG, J.; TANG, J.; LI, J. Expert Finding in a Social Network. In: KOTAGIRI, R. et al. (Ed.). *Advances in Databases: Concepts, Systems and Applications*. Springer Berlin Heidelberg, 2007, (Lecture Notes in Computer Science, 4443). p. 1066–1069. ISBN 978-3-540-71702-7 978-3-540-71703-4. Disponível em: http://link.springer.com/chapter/10.1007/978-3-540-71703-4_106. Citado 6 vezes nas páginas 15, 17, 26, 35, 36 e 43.

ZHANG, L. et al. Mechanism design for finding experts using locally constructed social referral web. In: *INFOCOM, 2012 Proceedings IEEE*. [S.l.: s.n.], 2012. p. 2896–2900. Citado 4 vezes nas páginas 21, 26, 30 e 38.

ZHAO, Z. et al. Expert Finding for Question Answering via Graph Regularized Matrix Completion. *Knowledge and Data Engineering, IEEE Transactions on*, v. 27, n. 4, p. 993–1004, abr. 2015. ISSN 1041-4347. Citado 3 vezes nas páginas 16, 34 e 36.